UMA ANÁLISE DE FATORES QUE INFLUENCIAM INTERAÇÕES ENTRE USUÁRIOS DO TWITTER

GIOVANNI VENTORIM COMARELA

UMA ANÁLISE DE FATORES QUE INFLUENCIAM INTERAÇÕES ENTRE USUÁRIOS DO TWITTER

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais – Departamento de Ciência da Computação como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Virgílio Augusto Fernandes Almeida

Belo Horizonte

Junho de 2012

Comarela, Giovanni Ventorim
C728a Uma análise de fatores que influenciam interações entre usuários do Twitter / Giovanni Ventorim Comarela. — Belo Horizonte, 2012
xxii, 64 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais – Departamento de Ciência da Computação
Orientador: Virgílio Augusto Fernandes Almeida

1. Computação - Teses. 2. Redes sociais on-line – Teses. I. Orientador. II. Título

519.6*04(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Uma análise de fatores que influenciam interações entre usuários do twitter

GIOVANNI VENTORIM COMARELA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. VIRGÍLIO AUGUSTO FERNANDES ALMEIDA - Orientador
Departamento de Ciência da Computação - UFMG

PROF. ADRIANO ALONSO VELOSO
Departamento de Ciência da Computação - UFMG

PROF. ARTUR ZIVIANI
Laboratório Nacional de Computação Científica - CNPq

Belo Horizonte, 01 de junho de 2012.

Eu dedico este trabalho ao meu pai (in memoriam) e a minha mãe.

Agradecimentos

Eu agradeço a Deus.

Eu agradeço aos meus pais. Foram eles que lutaram com dificuldades e me deram a chance de estudar.

Agradecimento especial ao meu orientador, professor Virgílio, por tudo. Pela paciência, pelo conhecimento transmitido e pelas oportunidades que me propiciou. Considero uma honra ter sido orientado por ele.

Obrigado aos professores Fabrício Benevenuto e Mark Crovella por toda a ajuda que me deram para que esse trabalho pudesse se realizar.

A todos os demais professores do DCC-UFMG com os quais tive contato e me ajudaram durante o mestrado.

A todos os membros do laboratório CAMPs. Pela amizade, companheirismo e ajuda, nos momentos fáceis e difíceis.

Aos vários amigos de república que tive durante esses mais de dois anos: Vitor, Rodolfo, Tiago, Flávio, Fabricio e Daniel. A ajuda não foi direta, mas o momentos de descontração foram cruciais.

Por último, mas não menos importante, eu agradeço ao CNPq pelo suporte financeiro, sem o qual definitivamente não poderia ter cursado o mestrado.

"O único lugar onde o sucesso vem antes do trabalho é no dicionário"
(Albert Einstein)

Resumo

Nesta dissertação estuda-se o problema de entender interações entre usuários na rede de informação Twitter. O problema é abordado em duas etapas: primeiro, é realizada uma caracterização extensiva de uma grande coleção de dados, a qual inclui todos os usuários, relações sociais e mensagens postadas na rede desde o início do serviço até julho de 2009. Os estudos mostram evidências de sobrecarga de informação. Como exemplo, algumas vezes os usuários passam por centenas de mensagens até encontrarem alguma que tem interesse em interagir. Estes resultados motivam a identificação de fatores que influenciam as probabilidades de respostas e compartilhamento de mensagens, tais como: interações passadas, a taxa de postagem de mensagem de quem as envia, a idade da mensagem e alguns elementos textuais que podem nela podem estar inseridos. Na segunda etapa, mostra-se que alguns destes fatores podem ser utilizados para melhorar o mecanismo de apresentação de mensagens para os usuários. Para isso, é construído um simples modelo para identificar períodos de atividades dos usuários ao longo do tempo. Após isso, este modelo é combinado com técnicas de aprendizado de máquina com o intuito de ordenar mensagens de acordo com suas respectivas probabilidades de interação. Através de estudos de simulação mostra-se que a fração de mensagens respondidas e compartilhadas próximas ao topo da lista de mensagens dos usuários cresce em até 60%.

Palavras-chave: Redes Sociais *Online*, Twitter, Interações.

Abstract

In information networks where users send messages to one another, the issue of information overload naturally arises: which are the most important messages? In this work we study the problem of understanding the importance of messages in Twitter. We approach this problem in two stages. First, we perform an extensive characterization of a very large Twitter data set which includes all users, social relations, and messages posted from the beginning of the service up to August 2009. We show evidence that information overload is present: users sometimes have to search through hundreds of messages to find those that are interesting to reply or retweet. We then identify factors that influence user response or retweet probability: previous responses to the same tweeter, the tweeter's sending rate, the age and some basic text elements of the tweet. In our second stage, we show that some of these factors can be used to improve the ordering of tweets as presented to the user. First, by inspecting user activity over time, we construct a simple on-off model of user behavior that allows us to infer when a user is actively using Twitter. Then, we explore two methods from machine learning for ranking tweets: a Naive Bayes predictor and a Support Vector Machine classifier. We show that it is possible to reorder tweets to increase the fraction of replied or retweeted messages appearing in the first positions of the list by as much as 60%.

Keywords: Online Social Networks, Twitter, Interactions.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A popularização da Internet ocorrida nos últimos anos fez com que as pessoas pudessem tirar maior proveito da *Web* para várias atividades. Entre os serviços mais utilizados estão as Redes Sociais *Online*, algumas contando com centenas de milhões de usuários ativos. Dentre estas redes pode-se citar o Twitter[1], a qual emergiu nos últimos anos como um serviço de *micro-blogging* e hoje se tornou uma vasta rede, hospedando um imenso fluxo de informação passando através de seus usuários.

Basicamente, utilizando o Twitter, os usuários podem "seguir" ou serem "seguidos" por outros sem a necessidade de reciprocidade. Esta relação implica que o "seguidor" receberá todas as mensagens (chamadas de *tweets*) postadas pelo "seguido". Uma marca importante do serviço é que as mensagens não podem ser compostas por mais do que 140 caracteres. São práticas comuns no Twitter compartilhar (originando *retweets*) e responder (originando *replies*) mensagens de outros usuários. Este simples arcabouço conquistou uma multidão de usuários, sendo que atualmente é raro visitar um *site* que não contenha a opção "compartilhar no Twitter".

A popularidade deste sistema pode ser traduzida em números. Ao completar 6 anos em março de 2012, a rede possuía cerca de 140 milhões de usuários ativos[2], os quais eram responsáveis por cerca de 340 milhões de mensagens por dia, o que significa que o Twitter lidava, naquela época, com cerca de um bilhão de mensagens a cada três dias. Como consequência, os usuários são sobrecarregados com uma grande quantidade de informação que podem não serem capazes de processar. Por exemplo, de acordo com Bernstein et al. [2010] um usuário ativo pode facilmente receber mais do que 1000 *tweets* por dia. Desta forma, torna-se difícil, ou até mesmo impossível conseguir absorver toda a informação recebida. No entanto, muitas mensagens são supérfluas,

[1] `http://www.twitter.com`

[2] `http://blog.twitter.com/2012/03/twitter-turns-six.html`

difíceis de entender ou até mesmo sem contexto, como é apresentado pelo recente estudo de André et al. [2012], o qual mostra que apenas 36% dos *tweets* recebidos valem a pena de serem lidos.

Com o exposto no parágrafo anterior pode-se perceber claramente a necessidade de estudos que possam ajudar os usuários do Twitter a separar o "joio do trigo". Para fazer isso, um primeiro passo é entender como os usuários "consomem" as mensagens que recebem e como interagem com outros usuários na rede. Pesquisas sobre o Twitter são mais recentes do que, por exemplo, em *e-mail* e *Web* e por consequência o conhecimento atual sobre este novo sistema não é tão sólido como nos demais casos. Desta forma, ainda existe um conjunto fundamental de questões a serem abordadas, entre as quais podem ser citadas: como usuários lidam com o fluxo de informação que recebem no Twitter? Qual o valor que os usuários atribuem as mensagens, em termos de como eles interagem com elas (respondendo ou compartilhando)? Quais fatores influenciam o fato de que um usuário irá ou não responder ou compartilhar uma dada mensagem? Responder estas questões pode ajudar no processo de gerenciamento para distinguir as mensagens mais importantes e interessantes.

Como motivação para estudar as perguntas estabelecidas no parágrafo anterior pode-se citar o fato de que caracterizar padrões de interações entre usuários e sistemas de computação permite identificar pontos de melhoria para o sistema. Em especial, no caso de Redes Sociais *Online* isto pode ser crucial uma vez que o modelo de negócio criado por elas implica em perda de lucro em caso de perdas significativas de usuários.

Esta dissertação está inserida justamente neste contexto, onde o objetivo é prover respostas para estas perguntas e também apresentar como utilizar as respectivas respostas em benefício dos usuários do Twitter. A seguir, apresenta-se o escopo da dissertação, enfatizando as metas a serem atingidas e os passos a serem seguidos. Por fim, a Seção 1.2 mostra como o restante do trabalho está organizado.

1.1 Escopo da dissertação

Este trabalho se incia com uma extensiva caracterização do comportamento dos usuários do Twitter. Para entender este comportamento, analisou-se uma grande coleção de dado que contém um histórico quase completo de todas mensagens trocadas entre os usuários por um período superior a 3 anos. O foco é dirigido ao entendimento e melhoria da forma com que os usuários interagem com as mensagens que recebem. Tipicamente, quando um usuário abre sua página Twitter ele se depara com um fluxo de mensagens que são apresentadas de maneira cronológica reversa, ou seja, as mais novas

ficam no topo da lista. Os usuários então podem ler e interagir com suas mensagens, de forma que esta última ação significa responder ou compartilhar um *tweet*. Assim, as mensagens escolhidas para serem respondidas ou compartilhadas proveem um indicador de quão "interessantes" elas são. Com base nesta premissa, as informações contidas nos dados coletados foram utilizadas para reconstruir os fluxos de mensagens dos usuários no intuito de responder as perguntas anteriormente estabelecidas como problema de pesquisa.

O primeiro conjunto de resultados é uma série de caracterizações relacionadas aos padrões de *replies* e *retweets*. De início, são apresentados indícios de que os usuários do Twitter estão expostos à sobrecarga de informação, mostrando que muitas vezes eles buscam por centenas de mensagens até encontrarem uma que têm interesse para interagir. Com esta motivação, são identificados fatores que influenciam nas probabilidades de interações, tais como: interações passadas com o mesmo usuário, a taxa de envio de mensagens de um usuário que postou determinado *tweet*, a idade da mensagem, o número de caracteres do texto e a presença de alguns elementos textuais comuns no vocabulário do Twitter.

Estes resultados não apenas revelaram alguns aspectos únicos do comportamento dos usuários do Twitter, como também motivaram a proposta de um método para modificar o mecanismo de apresentação de suas mensagens. Inspecionando a atividade dos usuários ao longo do tempo construiu-se um simples modelo para inferir quando estes estão ativamente usando a rede. Então, foram utilizados algoritmos de aprendizado de máquina para modificar a lista de mensagens dos usuários de forma a trazer as mais "interessantes" para o topo. Foram utilizados dois algoritmos de classificação: *Naive Bayes* e *Support Vector Machines*. Através de estudos de simulação mostrou-se que é possível reordenar os *tweets* de forma a aumentar a fração de *replies* e *retweets* nas p primeiras posições da lista de mensagens em até 60%. Tais resultados indicam que a metodologia proposta pode originar uma interface alternativa interessante para os usuários do Twitter.

1.2 Organização da dissertação

O restante desta dissertação está organizado da seguinte forma:

- Capítulo 2: discute uma série de trabalhos relacionados e como esta dissertação se diferencia deles;
- Capítulo 3: apresenta a coleção de dados analisada bem como detalhes do volume de mensagens que contém e de propriedades do grafo social;

- Capítulo 4: mostra evidências quantitativas de que os usuários do Twitter têm o problema de sobrecarga de informação e apresenta caracterizações de uma série de fatores que influenciam taxas de interações;

- Capítulo 5: com base em algumas características discutidas no capítulo 4 é proposta uma metodologia de reorganização das listas de mensagens dos usuários para que as mais interessantes (com maior chance de interação) sejam apresentadas primeiro. Após isso, os métodos são avaliados através de estudos de simulação, mostrando melhorias significativas;

- Capítulo 6: apresenta as considerações finais da dissertação e enumera algumas possibilidades de extensões e trabalhos futuros;

- Apêndice A: apresenta uma breve explicação sobre distribuições de Leis de Potência. É aconselhado para o leitor que não conhece as definições básicas ou a interpretação gráfica destas distribuições em escala logarítmica;

- Apêndice B: apresenta o procedimento do cálculo do tamanho amostral para se estimar proporções. Este procedimento é utilizado no Capítulo 5.

Parte desta dissertação foi aceita para publicação em uma conferência na área de *Web* e Mídias Sociais [Comarela et al., 2012].

Capítulo 2

Trabalhos Relacionados

O objetivo deste capítulo é apresentar um conjunto de trabalhos relacionados com o tema abordado nesta dissertação. Para esta finalidade, o texto está dividido da seguinte forma: a Seção 2.1 apresenta um conjunto de trabalhos relacionados com o Twitter, mas não necessariamente com o tema aqui abordado. A ideia é mostrar como este sistema está chamando a atenção da comunidade acadêmica e em que áreas isto está ocorrendo. A Seção 2.2 por sua vez, apresenta trabalhos diretamente relacionados ao tema abordado. Por fim, na Seção 2.3 são apresentadas as discussões e considerações finais do capítulo.

2.1 Uma visão geral do Twitter

Nesta seção o objetivo é discutir uma série de trabalhos relacionados com o Twitter. É importante ressaltar que não está no escopo deste texto apresentar todos os trabalhos em todas as áreas do conhecimento que tem relação com o Twitter, mas sim, alguns relevantes que foram estudados durante a elaboração desta dissertação e que de alguma forma contribuíram para sua realização.

Java et al. [2007] realizaram um dos primeiros estudos com intuito de caracterizar o Twitter ainda no início do serviço. Para uma coleção de dados de aproximadamente de 75 mil usuários e 1 milhão e 300 mil mensagens foram analisadas propriedades topológicas e geográficas da rede de informação. Além disso, identificou-se que os usuários usavam a rede para falar de suas atividades diárias e também com o objetivo de compartilhar informações. Uma análise de comunidades também foi realizada. Krishnamurthy et al. [2008] apresentam uma caracterização detalhada do Twitter para três coleções de dados distintas (certa de 100 mil usuários coletados no total). Foram identificadas classes distintas de usuários e seus respectivos comportamentos na

rede. Padrões de crescimento da rede foram analisados, assim como suas propriedades geográficas.

Kwak et al. [2010] apresentam uma caracterização do Twitter mais detalhada que os trabalhos anteriores e com uma coleção de dados consideravelmente mais rica. Foram coletados 41.7 milhões de perfis de usuários, 1.47 relações sociais e 106 milhões de mensagens postadas. Entre os principais achados sobre as propriedades topológicas podem ser citados: o fato de que as distribuições de graus que diferem de uma distribuição de Lei de Potência (como relatado pelos trabalhos citados anteriormente), diâmetro curto da rede e baixa reciprocidade. Uma análise inicial de influência foi conduzida comparando propriedades topológicas, como o *PageRank* [Brin & Page, 1998] com o total de mensagens compartilhadas que cada usuário teve. Além disso, foram analisados temas de grande popularidade na rede (conhecidos como *Trending Topics*) do ponto de vista de tópicos abordados e de difusão de informação.

Huberman et al. [2008] mostram que usuários do Twitter tem um número de "amigos" muito menor do que o número de pessoas que seguem na rede. A definição para o termo "amigos" de um usuário de pessoas que este usuário referenciou em uma mensagem pelo menos duas vezes. Em outras palavras, eles argumentam que, mesmo seguindo um grande número de pessoas, os usuários vão interagir com um número bem menor, sendo que estes, não necessariamente forma um subconjunto das pessoas que seguem na rede.

Com relação a estudos sobre a evolução do grafo social do Twitter ao longo do tempo podem ser citados os de: Romero & Kleinberg [2010], que introduzem o conceito de *fechamento triádico* para grafos direcionados e o utilizaram para um estudo a respeito de formação de relações sociais na rede; Yin et al. [2011] mostram que a maior parte das novas relações sociais são formadas entre usuários que estavam separados por dois graus na rede; Brzozowski & Romero [2011] analisam padrões estruturais da rede com intuito de avaliar a chance de formação de conexões; e Hopcroft et al. [2011] os quais apresentam uma metodologia para predizer a formação de relações recíprocas.

A forma que a informação se difunde entre os usuários do Twitter também é alvo de vários estudos. Entre trabalhos que tratam deste tema podem ser citados os de Rodrigues et al. [2011], Ienco et al. [2010], Galuba et al. [2010] e Romero et al. [2011]. Em especial, Weng et al. [2012] propõem um modelo para analisar o porque de algumas ideias se tornam populares e outras não. Pontos fortes da metodologia são os fatos de que o modelo é simples e não precisa considerar características externas à rede. Além disso, os autores analisam o fato de que os usuários tem "memória limitada" e por isso não tem como lembrar (e consequentemente participar) de todos os tópicos discutidos.

Relacionado ao tópico de difusão de informação está o de identificação de usuá-

rios chave (ou influentes) na rede. Este assunto foi amplamente discutido no contexto do Twitter uma vez que o conteúdo gerado por estes pode iniciar processos de cascata ou interessar a um grande público da rede. Existem na literatura vários artigos relacionados a este assunto. Exemplo são os de Cha et al. [2010], Weng et al. [2010], Bakshy et al. [2011] e Saez-Trumper et al. [2012].

Com a popularização do serviço também vieram uma variedade de entidades "maliciosas" utilizando o Twitter de forma automatizada e com finalidade distinta daquela para qual a rede foi idealizada. O trabalho de Benevenuto et al. [2010] apresenta um estudo de características deste tipo de usuários e as utiliza em algoritmos de aprendizado de máquina para encontrá-los.

Além destes, recentemente o Twitter também começou a ser utilizado para monitoração e previsão de determinados eventos. Exemplos deste tipo de aplicação são monitoramento de casos de dengue [Gomide et al., 2011], previsão de resultados de eleições [Saez-Trumper et al., 2011] e até monitoramento em tempo real de desastres naturais [Aljohani et al., 2011].

2.2 Análise de Interações

Existe uma grande quantidade de trabalhos na literatura que analisam e caracterizam a interação entre usuários e sistemas *Web*. Um exemplo recente é o trabalho de Radicchi [2009], o qual apresenta um estudo detalhado das distribuições de tempos entre eventos e de tempos de espera para três sistemas *Web* de naturezas distintas. Uma das principais conclusões deste estudo é a questão da dificuldade de se encontrar um modelo que seja útil para uma grande variedade de sistemas, preferencialmente com características heterogêneas. Apesar disso, uma constante em todos os achados é a presença de Leis de Potência[1] regendo as referidas distribuições. Este tipo de comportamento, intrínseco à natureza humana, de rajadas (*bursts*) e tem diversas explicações na literatura, principalmente no que tange interações com sistemas de computação, as quais fogem do escopo deste trabalho, mas podem ser encontradas nos trabalhos de Bárabási [2005], Vazquez et al. [2005], Brlanchard & Hongler [2007] Oliveira & Vazquez [2009], Crane et al. [2010] e Malmgren et al. [2008].

No contexto de Redes Sociais *Online*, interações também foram amplamente exploradas recentemente. Um estudo empírico bem detalhado é dado por Benevenuto

[1] Ver o Anexo A para uma introdução ao conceito de Leis de Potência.

[2010], o qual explora vários aspectos do Orkut[2], YouTube[3], Myspace[4], Hi5[5] e Linkedin[6].

Especificamente para o Twitter, além dos trabalhos já citados na Seção 2.1, pode-se mencionar o estudo de de Erramilli et al. [2011] que, além de várias caracterizações relacionada a atividade dos usuários do Twitter, apresenta um modelo de séries temporais para o total de mensagens postadas na rede diariamente. O trabalho em questão tem o objetivo de criar um "gerador" de carga para pessoas interessadas em questões de projeto de sistemas análogos ao Twitter.

Counts & Fisher [2011] utilizam técnicas de rastreamento de olhar para medir quais tipos de mensagens do Twitter recebem maior atenção dos usuários. Dentre os achados do trabalho, os autores mostram evidências de que mensagens que originam respostas refletem atenção e interesse dos usuários. Além disso, é apresentado que apenas mensagens acima de um certo limiar em termos de atenção e interesse são consideradas para serem compartilhadas. Uma desvantagem do trabalho em questão consiste no fato de ter feito uso de uma amostra de apenas 20 participantes. Embora este número seja pequeno ao ser comparado com os dos demais trabalhos citados, é importante mencionar que para a área de rastreamento de olhar, devido a restrições de tempo e custo, este é um número significativo.

No mesmo contexto, André et al. [2012] estudam o valor do conteúdo de uma mensagem. Através de estudos qualitativos os autores encontram que 36% de todas as mensagens postadas na rede valem a pena de serem lidas, 39% estão em uma zona neutra e 25% definitivamente não merecem atenção. Esses resultados indicam que os usuários toleram uma grande quantidade de informações não desejadas, fazendo com que mensagens realmente importantes, e potencialmente candidatas para interações, acabem sendo perdidas.

Claramente estes estudos motivam o projeto de ferramentas que auxiliem os usuários do Twitter a lidarem com a inundação de informação que recebem diariamente, facilitando o acesso à mensagens que possam ser mais "interessantes". Das Sarma et al. [2010] comparam mecanismos de ordenação de mensagens que fazem uso de avaliações de usuários em sistemas semelhantes ao Twitter. A métrica estudada pelos autores é acurácia versus custo, onde o custo é dado pelo número de revisões realizadas por mensagem. Através de modelos matemáticos e experimentos reais é mostrado que mecanismos binários de avaliação (*thumbs up-down ratings*) necessitam de um número

[2] www.orkut.com
[3] www.youtube.com
[4] www.myspace.com
[5] www.hi5.com
[6] www.linkedin.com

muito grande de revisões para gerar uma ordenação precisa, ao passo que a comparação de mensagens em pares diminui este problema.

No intuito de encontrar mensagens com maior chance de serem compartilhadas, Suh et al. [2010] estudam uma série de características com intuito de utilizar um modelo de regressão para dizer se uma dada mensagem será ou não compartilhada. No mesmo contexto, Hong et al. [2011] fazem uso de técnicas de aprendizado de máquina para prever a popularidade de mensagens em termos de quantas vezes elas serão compartilhadas. Uma característica comum destes dois trabalhos é o fato de que se baseiam em amostras coletadas do conjunto de todas as mensagens postadas no Twitter num determinado períodos de tempo. Além disso, os métodos propostos se baseiam em características das mensagens e dos usuários que as postam, de forma que não levam em consideração os usuários que recebem as mensagens.

Bernstein et al. [2010] abordam o problema de sobrecarga de informação propondo uma nova interface do Twitter. Utilizando uma técnica conhecida como LDA (*Latent Dirichlet Allocation*) [Blei et al., 2003], os autores agrupam as mensagens recebidas de acordo com tópicos, de forma que o usuário tenha mais facilidade em encontrar o conteúdo de interesse. Através de um protótipo construído e um estudo com usuários é mostrado que tal método propicia uma navegação mais fácil, amigável e que ajuda significativamente no processo de aliviar a sobrecarga de informação sofrida pelos usuários do Twitter.

2.3 Discussões

Comparado com este corpo de trabalhos relacionados, esta dissertação toma direções distintas, uma vez que o principal interesse é o estudo de interações dos usuários com as mensagens que recebem dos usuários que desejam seguir na rede, ou seja, não do fluxo total de mensagens geradas pelo Twitter ou em apenas um determinado tema. Além disso, tem-se interesse em melhorar a organização das mensagens recebidas pelos usuários. Como já dito, este tema já foi abordado por Bernstein et al. [2010], no entanto, o objetivo neste trabalho é ter uma metodologia simples que possa ser utilizada em dispositivos portáteis (em geral, identificação de tópicos é uma tarefa intensiva de CPU).

Por fim, é importante ressaltar que a grande maioria dos trabalhos relacionados está concentrado no estudo de *retweets* (mensagens compartilhadas), ao passo que nesta dissertação também é dada atenção aos *replies* (mensagens respondidas), mensagens as quais são comuns na rede mas bem menos exploradas na literatura.

Capítulo 3

Definições e Descrição da Coleção de Dados

O objetivo deste capítulo é apresentar um conjunto de definições e caracterizações básicas importantes para os próximos capítulos da dissertação. Primeiro, a Seção 3.1 apresenta definições importantes para o texto. A coleção de dados utilizada é descrita na Seção 3.2. Além disso, nesta seção mostra-se um conjunto de medições relacionadas ao volume de mensagens e ao grafo social do Twitter. Por fim, a Seção 3.3 realiza as discussões do capítulo.

3.1 Definições básicas

Nesta seção serão apresentados termos importantes no decorrer deste trabalho. No entanto, é importante salientar que não é o objetivo deste texto detalhar o funcionamento do Twitter e nem como os usuários interagem com tal sistema. Para o leitor não familiarizado aconselha-se uma visita ao *site* `www.twitter.com` ou a leitura do trabalho de Kwak et al. [2010].

Basicamente, o Twitter é uma plataforma de *micro-blogging* que permite atualmente que milhões de usuários postem milhões de mensagens (de até 140 caracteres) diariamente. Os usuários podem "seguir" pessoas, formando uma relação social direcionada. Essa relação social permite que as mensagens, conhecidas como *tweets*, postadas sejam recebidas pelos seguidores. O conjunto de todas as mensagens que uma pessoa recebe, de todas as pessoas que segue, é denominado *timeline*[1]. Existem dois tipos especiais de *tweets*:

[1] `http://support.twitter.com/articles/164083-what-is-a-timeline`

- *reply*: é uma mensagem direcionada a um usuário específico em resposta a uma de suas mensagens previamente postadas;
- *retweet*: este tipo de mensagem representa um compartilhamento de conteúdo postado por outro usuário;

Apesar de não pertencerem a língua portuguesa, as palavras *tweet*, *reply*, *retweet* e *timeline* não serão estilizadas em itálico no restante deste texto. Esta decisão foi tomada por razões puramente estéticas, uma vez que estes termos ocorrem com muita frequência no decorrer do trabalho. Além disso, convenciona-se o uso do verbo *compartilhar* para a ação de gerar um *retweet* e do verbo *responder* para a ação de gerar um *reply*.

3.2 Descrição do Conjunto de dados

A coleção de dados utilizada nesta dissertação é a mesma que apresentada no trabalho de Cha et al. [2010]. Estes dados consistem de um retrato do grafo social do Twitter em julho de 2009, um histórico quase completo de todos suas mensagens postadas desde sua criação em 2006 até Julho de 2009 e informações sobre os usuários. A seguir mais detalhes sobre o volume de dados e informações disponíveis[1] (apenas as mais relevantes neste trabalho):

- *Usuários*: foram coletados **54,981,152** usuários. Para cada um destes tem-se identificador numérico, *screen name* (apelido), número de seguidores, quantos usuários segue e data de criação da conta no Twitter;
- *Mensagens*: um total de **1,755,925,520** mensagens. Para cada uma tem-se identificador numérico, identificador do usuário que a postou, data (com resolução de segundos) e texto da mensagem. Além disso, se a mensagem for uma resposta para outra (de outro usuário) os dados contém os identificadores da mensagem e do usuário respondido. O mesmo não acontece para o caso dos retweets (detalhes na Seção 3.2.1);
- *Relações Sociais*: entre os usuários coletados foram identificadas **1,963,263,821** relações sociais. Estas relações são estabelecidas quando um usuário decide seguir outro. Para cada uma destas os dados contém o identificador numérico dos dois usuários envolvidos.

[1] O grafo social está disponível para *download* em `http://twitter.mpi-sws.org/data-icwsm2010.html`. Mensagens e Informações sobre usuários não estão publicamente disponíveis. Para ter acesso a estas, é necessário entrar em contato com um dos autores do trabalho referido previamente.

É importante ressaltar que esta coleção de dados é adequada ao trabalho proposto uma vez que contém um histórico praticamente completo de mensagens postadas por todos os usuários coletados. Esta característica é importante pois possibilita, junto com o grafo de relações sociais, obter aproximadamente o conjunto de mensagens recebidas por cada usuário (ou seja, seu respectivo timeline).

3.2.1 Identificação de retweets

Ao contrário dos replies, a coleção de dados analisada neste trabalho não contém informações sobre quais tweets são retweets e consequentemente a quais usuários e tweets se referem. Uma vez que estas informações são de extrema importância para tratar as questões estabelecidas, a seguinte heurística foi criada para contornar este problema: primeiro, para identificar os retweets fez-se uso do padrão utilizado pelos usuários. Neste padrão sempre estão presentes `RT @apelido` ou `via @apelido` junto a mensagem original, onde `apelido` é o *screen name* (único para cada usuário) do usuário no Twitter. Com isso, foi possível dizer quais tweets eram retweets e também qual foi o usuário que originalmente o postou.

Um segundo passo foi necessário para identificar qual tweet foi a origem de cada retweet. Para realizar esta tarefa, comparou-se cada retweet m com os tweets do usuário citado em m. Em mais detalhes, seja m_1 um tweet postado pelo usuário u_1 no tempo t_1 e m_2 um tweet postado pelo usuário u_2 no tempo t_2. Define-se neste trabalho que m_2 é um retweet de m_1 se as seguintes condições forem satisfeitas:

1. $u_1 \neq u_2$;
2. $t_2 > t_1$;
3. u_1 é citado em m_2 através de um dos padrões `RT @apelido` ou `via @apelido`;
4. Os tweets m_1 e m_2 são altamente similares. Para analisar esta condição, decidiu-se utilizar o índice de Jaccard [Baeza-Yates & Ribeiro-Neto, 2011] entre os conteúdos textuais de m_1 e m_2 e então verificar se seu valor é superior a um limiar ϵ. Desta forma, a similaridade entre estes tweets é definida por:

$$J(m_1, m_2) = \frac{|W_1 \cap W_2|}{|W_1 \cup W_2|}, \tag{3.1}$$

onde W_1 e W_2 são respectivamente os conjuntos das palavras que formam os textos de m_1 e m_2.

A heurística acima foi aplicada na coleção de dados analisada. Foram identificados aproximadamente 40 milhões de retweets, dos quais para 20 milhões foi possível identificar o usuário e a mensagem original compartilhada. É importante ressaltar que nesta metodologia é possível que para cada retweet exista mais de um tweet do mesmo usuário que satisfaça as quatro condições estabelecidas acima. Neste caso sempre foi escolhida a mensagem com o maior índice de Jaccard com o retweet em questão. Além disso, adotou-se para o limiar ϵ um valor de 0.2.

3.2.2 Notação

No intuito de formalizar as referências à coleção de dados são feitas as seguintes definições:

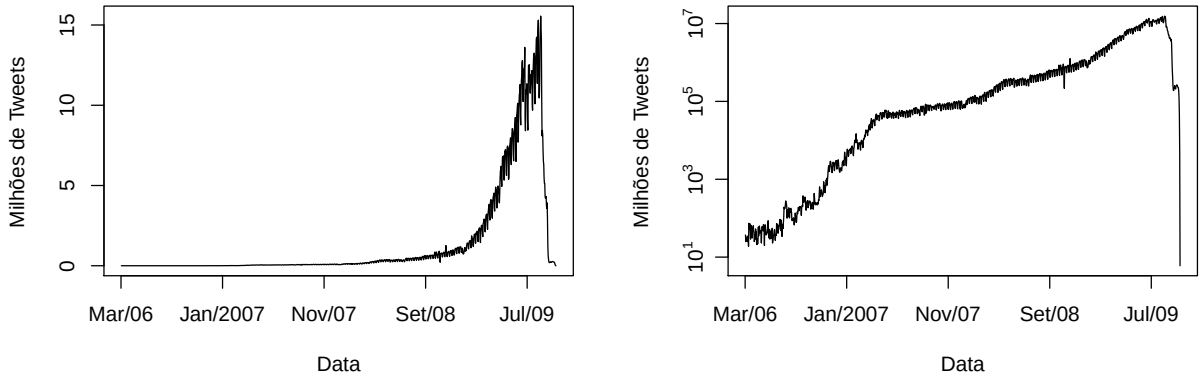
- U: conjunto de todos os usuários coletados;
- n_i: total de mensagens postadas pelo usuário i, $i \in U$;
- m_{i_j}: j-ésima mensagem postada pelo usuário i, $i \in U$ e $j = 1, \ldots, n_i$;
- M_i: conjunto de todas as mensagens postadas pelo usuário i, $i \in U$;
- In_i: conjunto de usuários que seguem o usuário i, $i \in U$;
- Out_i: conjunto de usuários que i segue, $i \in U$;
- M_{Out_i}: conjunto de todas as mensagens postadas pelos usuários que i segue, $i \in U$;
- TL_i: timeline do usuário i, ou seja, todas as mensagens postadas pelos usuários que i segue em ordem cronológica reversa, $i \in U$;
- $TL_i(t)$: estado de TL_i no instante de tempo t, ou seja, todas as mensagens que i recebeu até o instante t.

3.2.3 Volume de mensagens

Nesta seção apresenta-se uma caracterização do volume de mensagens presentes na coleção de dados. Este estudo é dividido em duas partes: primeiro mostra-se a evolução temporal da atividade dos usuários na rede. Segundo, analisa-se a distribuição do número de mensagens postadas por usuário.

3.2.3.1 Evolução Temporal

A Figura 3.1 apresenta o número de mensagens postadas diariamente que estão presentes nos dados coletados em duas versões: uma em escala linear (Figura 3.1a) e outra em escala logarítmica no eixo vertical (Figura 3.1b). Através destas figuras pode-se perceber um crescimento exponencial no período que compreende o início da rede até julho de 2009, chegando a picos de aproximadamente 15 milhões de mensagens por dia. É importante ressaltar que a quebra estrutural no final da curva ocorre devido a uma limitação do processo de coleta, uma vez que foi necessário aproximadamente um mês para obter todas as informações.



(a) Escala linear

(b) Escala logarítmica

Figura 3.1: Número de mensagens na coleção de dados (série temporal diária).

3.2.3.2 Mensagens por Usuário

A Figura 3.2 mostra a Função de Distribuição Acumulada Complementar (CCDF, do inglês *Complementary Cumulative Distribution Function*) do número total de mensagens postadas por cada usuário em U. De acordo com esta figura tem-se que a maior parte dos usuário postam poucas mensagens. Por exemplo, 90% de todas os usuários postaram 10 ou menos mensagens e 99% menos de mil. Além disso, pode-se observar que a cauda da distribuição se comporta de acordo com uma distribuição de Lei de Potência com expoente $\alpha = 2.9$ (o processo de regressão linear originou $R^2 = 0.991$)[2].

[2]O R^2 permite avaliar a qualidade de uma regressão linear. Quanto mais próximo de 1 for o valor, melhor é o resultado. Um ajuste ruim origina um R^2 próximo de 0. Ao leitor interessado, recomenda-se o trabalho de Jain [1991].

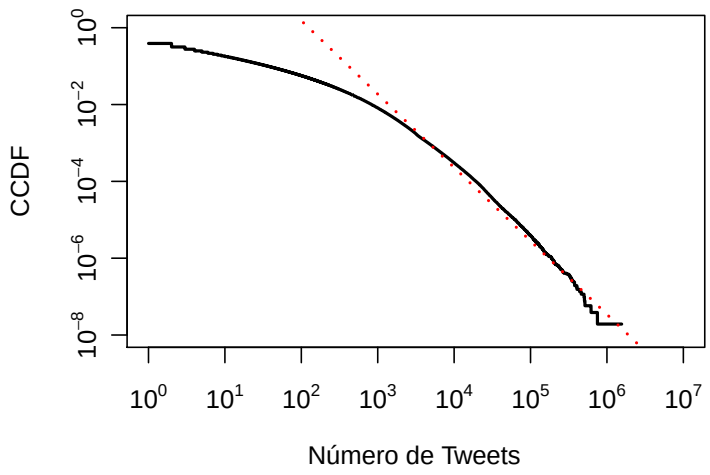


Figura 3.2: Distribuição do número de tweets por usuário.

3.2.4 Propriedades do grafo social

O objetivo desta seção é apresentar algumas características do grafo social do Twitter. Este estudo é importante para permitir uma melhor compreensão da coleção de dados em análise. Além disso, alguns dos resultados aqui apresentados não foram encontrados em outros trabalhos da literatura.

Nas análises seguintes a rede social será representada por um grafo $G(V, E)$, onde V é o conjunto de todos os usuários em U e E é um conjunto de pares ordenados (u, v) tais que $u, v \in V$, $u \neq v$ e se o usuário u segue v no Twitter então $(u, v) \in E$. Com esta definição tem-se que G é um grafo direcionado não ponderado e sem auto-*loops*. Além disso, define-se também a versão não direcionada de G por $G'(V', E')$, onde $V' = V$ e E' é tal que se $(u, v) \in E$ então $(u, v) \in E'$ e $(v, u) \in E'$.

3.2.4.1 Distribuições de graus

Uma das métricas mais simples e usuais para a análise de grafos sociais é a distribuição de graus dos nós (usuários). Uma vez que o grafo social do Twitter é direcionado, duas análises serão apresentadas: uma para os graus de entrada (*in-degree*) e outra para os graus de saída de cada nó (*out-degree*), representando respectivamente para um usuário u, o número de pessoas que seguem u e o número de pessoas que u segue na rede.

A Figura 3.3 apresenta a Função de Distribuição Acumulada Complementar (CCDF) para estas variáveis. Para a distribuição do *in-degree* pode-se perceber uma semelhança com uma curva que segue uma Lei de Potência. No entanto, é importante ressaltar que a cauda da distribuição foge a esse comportamento devido a um

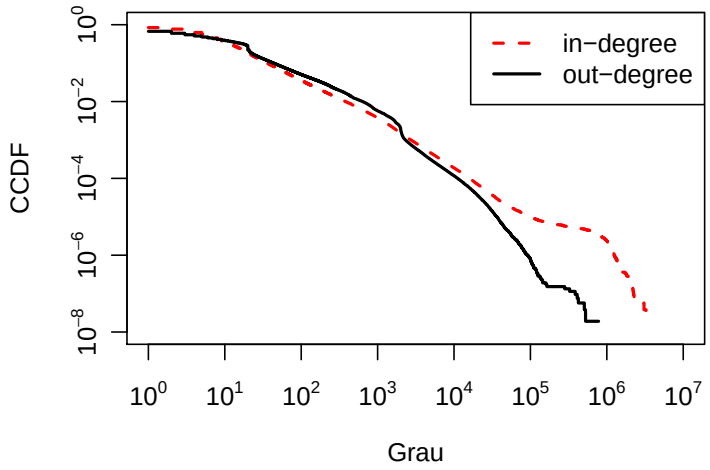


Figura 3.3: Distribuições de Graus.

grande número de usuários com muitos seguidores, em geral celebridade que usam a rede para manter contatos com seus fãs [Kwak et al., 2010]. Através de uma regressão linear simples obteve-se o expoente da distribuição de Lei de Potência $\alpha = 2.15$ (com $R^2 = 0.994$), para o intervalo de 10 a 10^5 no eixo horizontal.

A distribuição de *out-degree* também possui alguns pontos interessantes. Existem dois degraus na curva, um em 20 e outro em 2000. Isto ocorre uma vez que no início do Twitter havia uma recomendação de 20 pessoas a serem seguidas para os novos usuários e uma restrição de que ninguém na rede poderia seguir mais do que 2 mil pessoas. Atualmente nem a recomendação nem a restrição citadas existem mais. Além disso, pode-se perceber que o comportamento da curva no intervalo $[10, 2000]$ é diferente do que em $[2000, 100000]$. Por isso, decidiu-se procurar uma distribuição de Lei de Potência para cada caso. No primeiro, obteve-se $\alpha = 2.20$ ($R^2 = 0.972$) e no segundo $\alpha = 3.4$ ($R^2 = 0.998$). O ajuste para o intervalo completo também foi feito, originando $\alpha = 2.36$ ($R^2 = 0.968$).

É importante ressaltar que valores típicos para o expoente da distribuição de Lei de Potência estão tipicamente entre 1 e 3.5 no que se refere o contexto desta seção [Ebel et al., 2002].

3.2.4.2 Reciprocidade

Após a análise das distribuições de graus uma questão que surge naturalmente é: Quantas relações sociais são recíprocas? em outras palavras, do total de arestas em E, quantas são tais que se $(u, v) \in E$ então $(v, u) \in E$? No grafo G encontrou-se que de todos os pares ligados por pelo menos uma aresta apenas 17.78% são recíprocos. Este

resultado indica uma discrepância nas relações sociais, onde tornou-se comum usuários se interessarem pelo conteúdo postado por outros, no entanto, o interesse é não retribuído. Além disso, este número está de acordo com os achados de Kwak et al. [2010], o quais mostram que o Twitter tem uma baixa reciprocidade ao ser comparado com outras Redes Sociais *Online*.

Para uma análise mais profunda deste mérito decidiu-se olhar para esta quantidade de forma local. Para isso, foram utilizadas as seguintes medidas de reciprocidade definas por Mislove et al. [2007]:

$$R_{Out}(v) = \frac{|In_v \cap Out_v|}{|Out_v|} \tag{3.2}$$

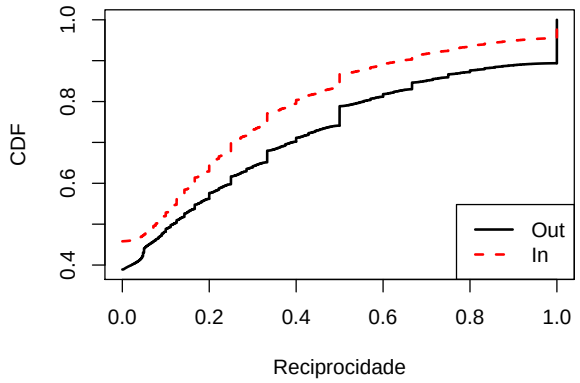
e

$$R_{In}(v) = \frac{|In_v \cap Out_v|}{|In_v|}. \tag{3.3}$$

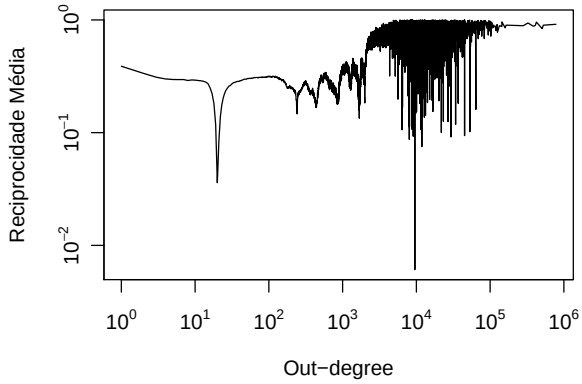
Basicamente, com estas definições tem-se que $R_{Out}(v)$ quantifica a reciprocidade local de v em relação aos usuários da rede que segue e $R_{In}(v)$ o mesmo, mas em relação a quem segue v.

A Figura 3.4 apresenta uma análise de $R_{Out}(v)$ e $R_{In}(v)$ para todos os usuários em U tais que $Out_v > 0$ e $In_v > 0$ respectivamente. Em ambos os casos mais de 40 milhões de usuários foram utilizados. Através da Função de Distribuição Acumulada (CDF, do inglês *Cumulative Distribution Function*) destas variáveis (Figura 3.4a) tem-se que em geral a reciprocidade relacionada aos nós de saída são maiores que as dos nós de entrada. Este fato indica que dado um nó v, é mais comum v ser seguido por quem ele segue do que o caso contrário. Além disso, pode perceber que 80% dos valores são menores que 0.5 e 0.6 para $R_{In}(v)$ e $R_{Out}(v)$ respectivamente, o que realmente confirma a afirmação anterior de que o Twitter possui uma rede marcada por baixa reciprocidade.

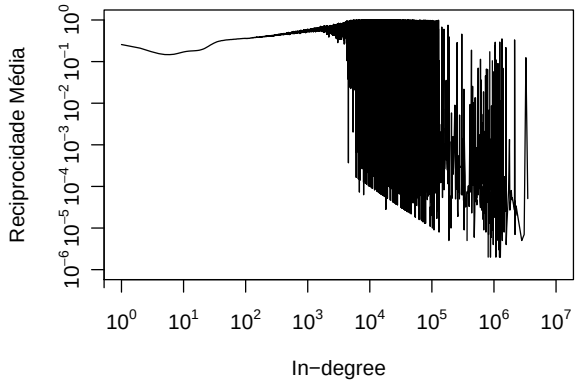
Outra análise realizada foi investigar qual a relação de $R_{Out}(v)$ com Out_v e de $R_{In}(v)$ com In_v. Para isso, calculou-se a reciprocidade média para cada grau e construiu-se um gráfico que relaciona estas duas quantidades em escala logarítmica para os dois eixos. No primeiro caso (Figura 3.4b) percebe-se que não existe nenhuma relação, positiva ou negativa, entre Out_v e $R_{Out}(v)$, no entanto, a curva possui alguns pontos interessantes para serem analisados. Um destes pontos ocorre próximo do valor 20 no eixo horizontal, o qual é uma queda brusca no valor da reciprocidade média. Este fato é facilmente explicado pela recomendação de usuários que o Twitter fazia para novos usuários da rede, como já citado na seção anterior. O próximo ponto é próximo ao valor mil, onde existe uma mudança da curva de uma linha para uma nu-



(a) Distribuições de $R_{Out}(v)$ e $R_{In}(v)$



(b) *in-degree* × R_{In} média



(c) *out-degree* × R_{Out} média

Figura 3.4: Análise da reciprocidade de G.

vem de pontos. Essa mudança é explicada pelo fato de que existem poucos usuários com *out-degree* maior que 1000, ou seja, devido ao baixo número de observações para o cálculo da média a variância dos resultados aumenta significativamente. Por fim, observa-se na cauda da curva que existem muitos pontos com alto *out-degree* e com reciprocidade próxima a 1.

Com relação a Figura 3.4c observa-se que existe uma relação positiva entre o *in-degree* dos usuários e $R_{In}(v)$. Em outras palavras, no caso geral, ao aumentar o número de pessoas que seguem um determinado nó v, também aumenta o número de pessoas que v segue neste grupo. No entanto, este comportamento parece ser válido apenas até quando os usuários tem aproximadamente 1000 seguidores. Após esse valor, a curva torna-se uma nuvem de pontos, fato que tem a mesma explicação apresentada para a Figura 3.4b. Além disso, é importante ressaltar que para valores extremamente altos de In_v a nuvem de pontos parece decrescer, indicando que pessoas com alto *in-degree*, em geral celebridades, tendem a não retribuir a grande massa de fãs que as seguem.

3.2.4.3 Graus de separação

Esta seção apresenta uma análise do comprimento do caminho mínimo entre dois usuários do grafo social do Twitter. Uma análise exata exigiria computar o caminho mínimo de todos os vértices para todos os vértices de G. Devido ao alto custo computacional que esta tarefa demandaria, optou-se em estimar a distribuição desta variável através do procedimento probabilístico usado por Ahn et al. [2007].

Foram extraídas amostras aleatórias de tamanho k de V, e para cada uma delas computou-se o comprimento do caminho mínimo de cada nó da amostra para todos os demais de V. O primeiro valor considerado para k foi 2000. Após isso, esta quantidade foi aumentada até 10000, sendo que ao atingir este valor a distribuição de probabilidade do comprimento dos caminhos mínimos praticamente não se alterava ao aumentar k novamente.

A Figura 3.5 apresenta a distribuição de probabilidade desta variável para $k = 10000$ considerando dois casos distintos, o grafo G e sua versão não direcionada G'. Em ambos os casos, tem-se que a moda da distribuição é 4 com uma média de 4.44 para G e de 3.68 para G'. Além disso, o diâmetro, maior valor encontrado para o valor do caminho mínimo, para G é 17 e para G' 16.

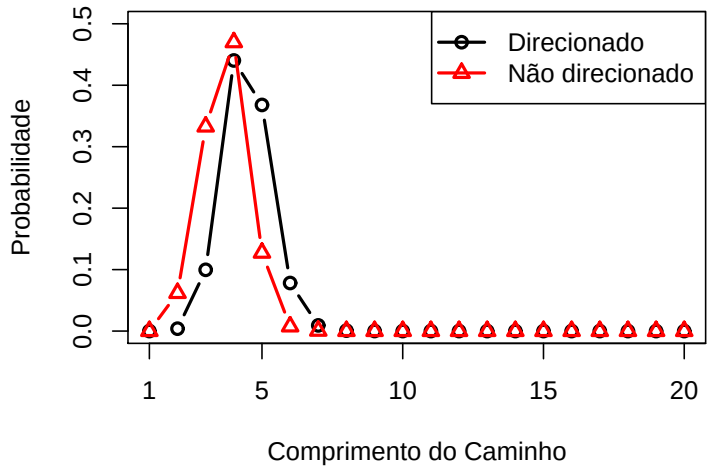


Figura 3.5: Estimativa da distribuição do comprimento do caminho mínimo de G.

É interessante notar que os valores apresentados no parágrafo anterior são relativamente menores que os reportados no famoso estudo de Milgram [1967], o qual reporta uma moda de 6. No entanto, são próximos quando comparandos com os de outra Rede Social *Online* popular, o Facebook[3], o qual tem um caminho mínimo médio

[3] `www.facebook.com`

de 4.74 [Backstrom et al., 2011].

3.2.4.4 Coeficiente de agrupamento

O coeficiente de agrupamento é uma medida local que indica o quão conexo é um grafo na vizinhança de um determinado nó. Dado um nó v, o coeficiente de agrupamento de v é a fração do número de arestas existentes em relação o número total de arestas que podem existir entre os vizinhos de v. No entanto, dado que G é um grafo direcionado, decidiu-se utilizar medidas que diferenciam as arestas de entrada e saída de um nó, gerando duas medidas distintas. Formalmente, tem-se que [Watts & Strogatz, 1998]:

$$CC_{In}(v) = \frac{|E_{In}(v)|}{|In_v|(|In_v| - 1)} \tag{3.4}$$

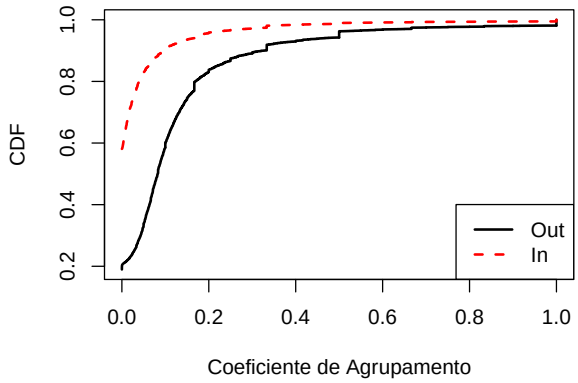
para o conjunto de vértices de entrada (que seguem v) e

$$CC_{Out}(v) = \frac{|E_{Out}(v)|}{|Out_v|(|Out_v| - 1)} \tag{3.5}$$

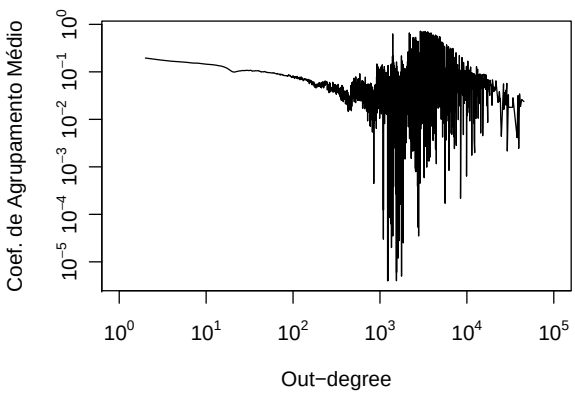
para o conjunto de vértices de saída (que v segue), onde $E_{In}(v) = \{(i,j) : i,j \in In_v \text{ e } (i,j) \in E\}$ e $E_{Out}(v) = \{(i,j) : i,j \in Out_v \text{ e } (i,j) \in E\}$. Em resumo, $CC_{In}(v)$ representa a fração de arestas formadas dos nós que seguem v e $CC_{Out}(v)$ a fração de arestas formadas entre os nós que v segue.

A Figura 3.6 apresenta uma análise de $CC_{In}(v)$ e $CC_{Out}(v)$ para uma amostra aleatória de usuários v em V (totalizando um milhão de usuários analisados) tais que $In_v > 1$ e $Out_v > 1$ respectivamente. Foi necessário se trabalhar com amostras devido ao custo computacional para se calcular o coeficiente de agrupamento para todos os nós de G. Através das CDFs destas variáveis (Figura 3.6a) percebe-se que o coeficiente de agrupamento relacionado aos nós de saída é maior que o dos nós de entrada. Este fato indica que dado um nó v é mais comum se formar (ou haver) uma aresta entre os nós que v segue, do que daqueles que seguem v. Além disso, pode-se perceber que cerca de 90% dos valores são menores que 0.1 e 0.4 para $CC_{In}(v)$ e $CC_{Out}(v)$ respectivamente, o que sugere que o grafo G é marcado por pequenos valores de coeficiente de agrupamento.

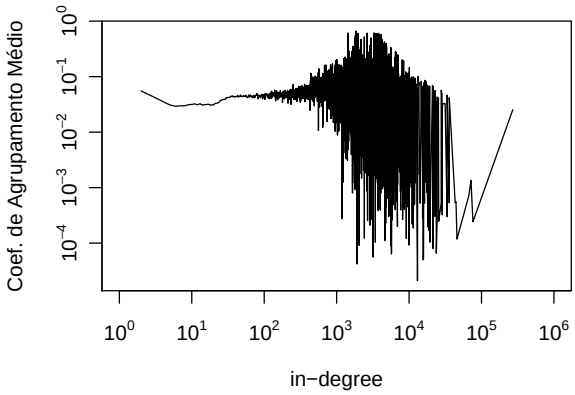
Outra análise realizada foi investigar a relação de $CC_{Out}(v)$ com Out_v e de $CC_{In}(v)$ com In_v. Para isso calculou-se o coeficiente de agrupamento médio para cada grau e construiu-se um gráfico que relaciona estas duas quantidades em escala logarítmica para os dois eixos. No primeiro caso (Figura 3.6b) percebe-se inicialmente uma relação negativa entre as duas variáveis, ou seja, a medida que o *out-degree* aumenta o valor médio de CC_{Out} diminui. No entanto, este comportamento só é válido até cerca de mil pessoas seguidas. Após isso, a curva se torna uma nuvem de pontos,



(a) Distribuições de $CC_{In}(v)$ e $CC_{Out}(v)$



(b) *Out-degree* × CC_{Out}



(c) *Out-degree* × CC_{In}

Figura 3.6: Análise do Coeficiente de Agrupamento de G.

onde são válidos os mesmos comentários feitos na Seção 3.2.4.2 (inclusive para a Figura 3.6c).

Com relação a Figura 3.6c pode-se perceber que não existe uma relação, positiva ou negativa, entre as variáveis. Este fato indica que fato de dois usuários seguirem um determinado nó v não contribui para a formação de um aresta entre eles, não importando que v tenha muitos ou poucos seguidores.

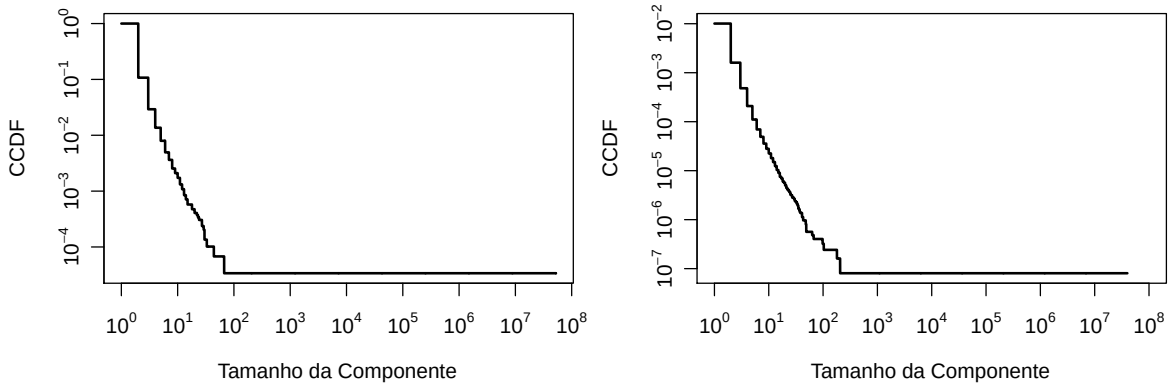
3.2.4.5 Componentes conexas

Em um grafo não direcionado um componente conexa é definida como um subconjunto C de V no qual sempre existe um caminho entre qualquer par de seus vértices. No caso de um grafo direcionado esta definição é estendida para dois conceitos diferentes. Diz-se haver uma componente *fortemente* conexa (SCC, do inglês *Strongly Connected Component*) quando o caminho em quesão é direcionado e uma componente *fracamente*

conexa (WCC, do inglês *Weakly Connected Component*) em caso contrário.

Foram investigados o número e o tamanho das SCCs e WCCs do grafo social do Twitter. Para o caso das WCCs é suficiente utilizar um procedimento de busca em largura com pequenas modificações tendo como entrada o grafo G', ou seja, a versão não direcionada de G. No caso das SCCs, o procedimento é um pouco mais elaborado, onde foram utilizados dois procedimentos de busca em profundidade assim como apresentado por Cormen et al. [2009].

A Figura 3.7 apresenta as distribuições de probabilidade para o tamanho de todas as SCCs e WCCs encontradas em G. Em ambos os casos pode-se perceber uma grande quantidade de pequenas componentes e a presença de apenas uma componente gigante. De fato, foram encontradas cerca de 30 mil componentes fracamente conexas, sendo a maior composta por mais de 52 milhões de nós. No caso das componentes fortemente conexas, foram identificadas cerca de 12 milhões, tendo a maior mais de 40 milhões de nós.



(a) Fracamente Conexas (b) Fortemente Conexas

Figura 3.7: Distribuições do tamanho das componentes conexas de G.

3.3 Discussões

Este capítulo apresentou definições importantes para o entendimento do restante da dissertação. Além do mais, foi apresentada uma descrição detalhada da base de dados analisada e quais passos foram necessários para que um tipo especial de mensagem, os retweets, pudessem ser identificadas.

Através dos estudos descritivos verificou-se que a distribuição do número de tweets por usuário segue uma distribuição de Lei de Potência. Além disso, mostrou-se que o número de tweets postados por dia cresceu exponencialmente durante o período coletado.

Quanto a estrutura da rede, observou-se que o grafo social do Twitter possui baixa reciprocidade, baixo coeficiente de agrupamento e distribuições de *in-degree* e *out-degree* que se comportam de acordo com uma lei de potência para a maior parte dos nós. Além disso, o grafo possui componentes forte e fracamente conexas que abrangem uma parte significativa dos nós, indicando desta forma que a presença de componentes gigantes.

Com o exposto, tem-se que a coleção de dados analisada possui uma rede com as estruturas básicas de uma rede social com um histórico significativo de mensagens trocadas entre os usuários, fazendo com que esta coleção seja ideal pra o propósito dos próximos capítulos.

Capítulo 4

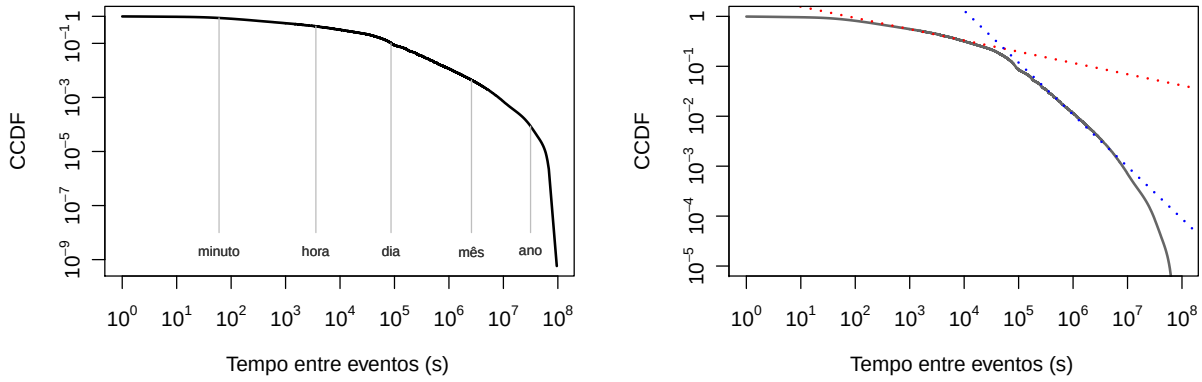
Caracterizações de atividades e interações

O objetivo deste capítulo é apresentar um conjunto de caracterizações que descrevem aspectos da atividade dos usuários do Twitter presentes na coleção de dados descrita no capítulo anterior. A meta principal destas caracterizações é identificar e entender fatores que afetam taxas de interações. Para este fim, o capítulo está dividido da seguinte maneira: a Seção 4.1 caracteriza o tempo entre eventos gerados pelos usuários no intuito de entender o quão intensamente estes postam mensagens na rede. A Seção 4.2 apresenta a distribuição empírica do tempo de espera para que tweets sejam respondidos ou compartilhados. Na Seção 4.3 apresenta-se uma discussão e caracterização de fatores que afetam taxas de replies e retweets na rede. Por fim, a Seção 4.4 discute os principais achados do capítulo.

4.1 Tempo entre eventos

O objetivo desta seção é entender o comportamento dos usuários do Twitter no que tange a intensidade com que eles interagem com a rede. Para este fim, decidiu-se analisar a distribuição global da variável aleatória relativa ao tempo entre eventos (tweets) gerados pelos usuários, a qual será denotada por Δt.

Suponha que as n_i mensagens do usuário i tenham ocorrido nos instantes de tempo $t_{i_1}, t_{i_2}, \ldots, t_{i_{n_i}}$, com $t_{i_1} < t_{i_2} < \cdots < t_{i_{n_i}}$. Com esta informação é possível calcular o tempo entre eventos consecutivos para cada usuário dados por $\Delta t_{i_1} = t_{i_2} - t_{i_1}, \ldots, \Delta t_{i_{n_i-1}} = t_{i_{n_i}} - t_{i_{n_i-1}}$. Estes valores foram computados para todos os usuários de U que possuem duas ou mais mensagens postadas (cerca de 20 milhões de usuários).



(a) CCDF de Δt

(b) Ajuste para Δt

Figura 4.1: Análise da distribuição de Δt.

Após esta etapa, estes resultados foram agregados para obter a distribuição empírica de Δt.

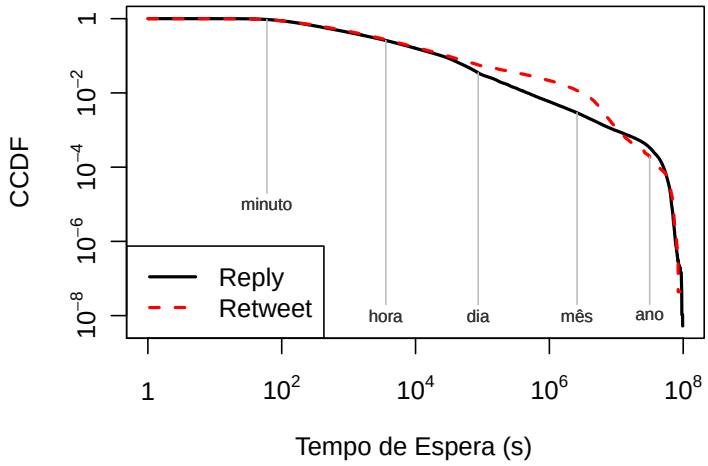
A Figura 4.1 apresenta duas versões da Função de Distribuição Acumulada Complementar (CCDF) empírica de Δt. Na primeira (Figura 4.1a), pode-se perceber uma grande heterogeneidade, uma vez que os valores de Δt variam de 1 segundo até aproximadamente 3 anos. Além disso, a curva possui 3 pontos que merecem especial atenção devido a mudanças que ocorrem em seu comportamento: *i)* perto de 1 minuto, tem-se que apenas 13% dos valores de Δt estão abaixo do referido valor. Isto indica que não é muito comum usuários postarem muitas mensagens em um intervalo muito curto de tempo; *ii)* próximo de um dia, onde tem-se $P(\Delta t \leq 1 \text{ dia}) = 0.9$. Este resultado indica que grande parte dos usuários analisados não conseguem ficar mais que um dia sem postar um tweet; e *iii)* perto de 10^8 segundos (aproximadamente 3 anos), valor que na época da coleta era a idade do Twitter.

A Figura 4.1b apresenta a mesma distribuição, no entanto com foco maior nos pontos 1 e 2 citados no parágrafo anterior. Isso foi feito com o objetivo de ter uma melhor visualização dos ajustes realizados. No primeiro intervalo considerado, de um minuto a um dia, obteve-se uma distribuição de Lei de Potência com $\alpha = 1.22$ ($R^2 = 0.984$) enquanto para o segundo, de um dia a um ano, obteve-se $\alpha = 2.04$ ($R^2 = 0.986$). Esses valores de expoentes confirmam o que foi exposto no parágrafo anterior sobre o fato que a distribuição de Δt possui várias mudanças de comportamento. Uma possível explicação para estas mudanças é existência de classes distintas de usuários no que se refere a intensidade de interação com o Twitter.

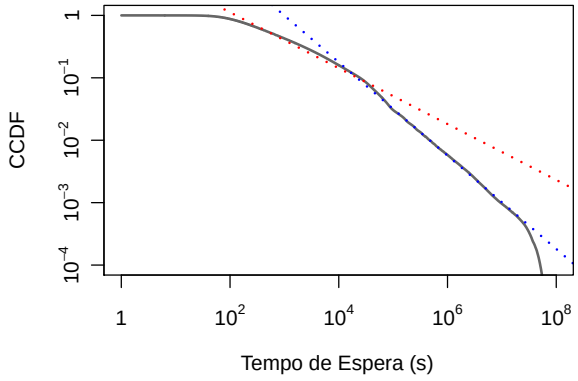
4.2 Tempo de Espera

Nesta seção analisa-se a distribuição do tempo que tweets esperam para serem respondidos ou compartilhados. Formalmente, tem-se interesse em analisar duas variáveis aleatórias: τ_p o tempo de espera para que uma mensagem seja respondida (um reply); e τ_t o tempo de espera para que uma mensagem seja compartilhada (um retweet). Para este fim, para todos replies e retweets da coleção de dados, calculou-se a diferença de tempo entre os instantes em que o tweet foi postado e que foi respondido (ou compartilhado).

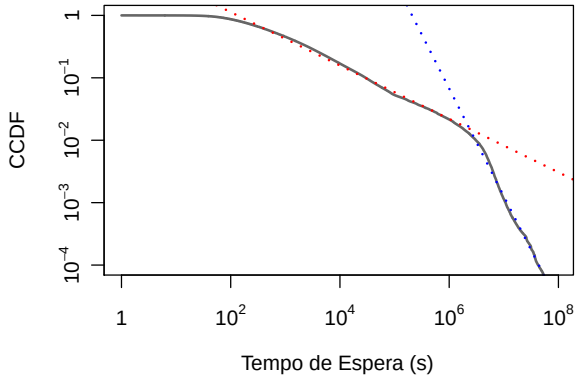
A Figura 4.2a apresenta as CCDFs de τ_p e τ_t. Através destas curvas pode-se perceber que alguns comentários feitos para a distribuição de Δt aqui também são válidos, principalmente no que tange os pontos de mudança de comportamento próximos



(a) Distribuições de τ_p e τ_t



(b) Ajuste para τ_p



(c) Ajuste para τ_t

Figura 4.2: Análise das distribuições de τ_p e τ_t.

a um minuto, um dia e após um ano. Além disso, tem-se que praticamente todas as mensagens esperam mais que um minuto até serem respondidas ou compartilhadas e que aproximadamente 90% delas esperam até dez mil segundos (aproximadamente 3 horas). Não obstante, existem mensagens que esperam períodos de tempo extremamente longos, chegando a anos. Estes resultados mostram evidências iniciais de que usuários do Twitter estão expostos a sobrecarga de informação fazendo com que muitas vezes demorem muito para interagir com mensagens de interesse.

Outra característica importante nesta figura é que as curvas de τ_p e τ_t se desencontram após 10^4s. Através desta mudança pode-se perceber que replies tendem a ocorrer mais rapidamente do que retweets, sugerindo que os usuários tem mais chance de compartilhar do que de responder mensagens antigas.

Para complementar estes resultados, as Figuras 4.2b e 4.2c apresentam os ajustes das distribuições de Lei de Potência para as caudas das curvas. Assim como ocorreu para Δt tem-se que dois regimes diferentes em cada caso foram identificados. A Tabela 4.1 apresenta os valores encontrados para os expoentes, onde pode-se perceber que os ajustes são adequados pela análise dos valores do R^2 da regressão linear. Além disso, percebe-se a proximidade dos valores de α referentes aos primeiros intervalos para replies e retweets, o que não se repete para os intervalos seguintes.

Tabela 4.1: Descrição dos ajustes distribuições de Lei de Potência para τ_p e τ_t

Interação	Reply (Figura 4.2b)		Retweet (Figura 4.2c)	
Intervalo	[1 minuto, 1 dia]	(1 dia, 1 ano]	[1 minuto, 1 mês]	(1 mês, 1 ano]
α	1.44	1.74	1.42	2.72
R^2	0.976	0.999	0.994	0.991

4.3 Caracterizações de Fatores que Influenciam Interações

Nesta seção serão apresentadas caracterizações relacionadas com a preferência do usuário no que tange interações com outras mensagens e usuários no Twitter. As características aqui exploradas são: a idade da mensagem, a ocorrência (ou não) de interações passadas, taxa de envio de mensagens e a presença de alguns atributos textuais comuns em tweets.

Para realizar cada uma dessas caracterizações, a reconstrução do timeline dos usuários é uma etapa a ser realizada. Em outras palavras, para cada usuário é necessário recuperar da coleção de dados os tweets de todas as pessoas que ele segue em ordem

cronológica reversa. Infelizmente, devido ao custo computacional (toda a coleção tem quase 1TB), não foi possível realizar esta tarefa para todos os usuários de U. Para contornar este problema decidiu-se utilizar amostras aleatórias de usuários.

Quatro amostras aleatórias de usuários foram extraídas, as quais serão denotadas por S_1, S_2, S_3 e S_4 e foram obtidas a partir do conjuntos de usuários que possuíam mais que 2, 10, 100 e 1000 mensagens postadas respectivamente. A ideia de considerar essa diferença é verificar se a atividade dos usuários tem algum impacto nas variáveis analisadas. Vale lembrar que no restante deste capítulo, quando nenhuma amostra é explicitamente relacionada a um resultado, tem-se que este é referente a toda a coleção de dados (todos os usuários).

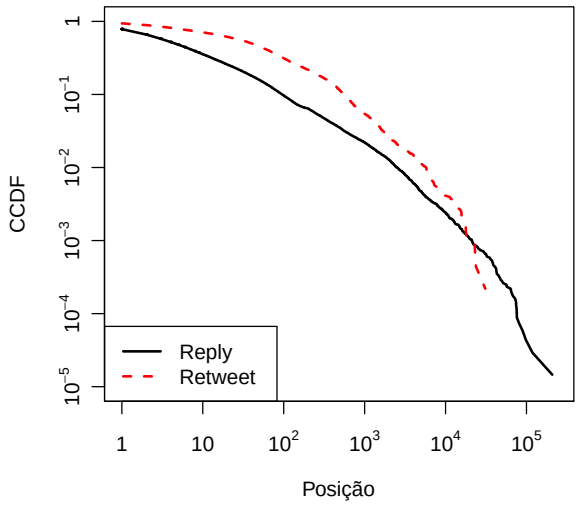
4.3.1 Idade da mensagem

Nesta seção tem-se o interesse em responder a seguinte pergunta: *tweets mais novos tem mais chances de serem respondidos e compartilhados?*

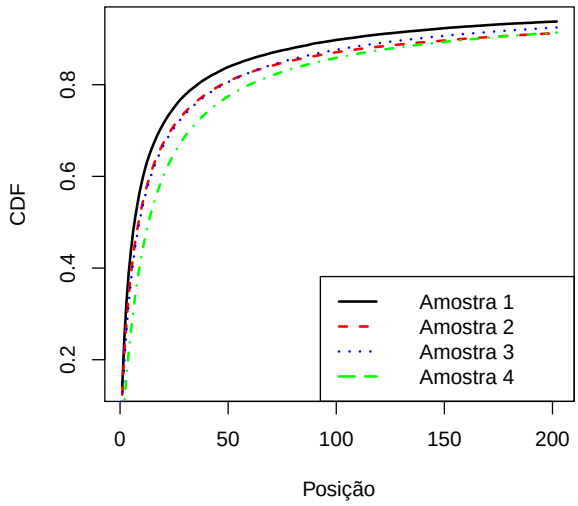
Para responder essa questão, decidiu-se não olhar especificamente para a idade da mensagem medida em unidades de tempo, mas sim para a posição no timeline em que ela se encontrava num determinado instante de interesse. Convenciona-se que a mensagem mais nova do timeline, a que está no topo, está na posição 0 do timeline, a próxima na posição 1 e assim por adiante. Salientando que essa decisão foi tomada por duas razões: *i)* uma vez o Twitter mostra mensagens para os usuários em ordem cronológica reversa existe uma relação direta entre sua posição no timeline e idade; e *ii)* trabalhar com variáveis discretas é mais simples neste contexto. Desta forma, dado um usuário i, para cada $j = 1, \ldots, n_i$, se m_{i_j} é um reply (ou retweet) busca-se em $TL_i(t_{i_j})$, o timeline de i quando m_{i_j} é postada, pela mensagem que foi respondida (compartilhada). Com isso, é possível identificar em que posição de $TL_i(t_{i_j})$ m_{i_j} estava. Esse procedimento foi repetido para todo $i \in S_k$, $k = 1, ..., 4$. Por simplicidade, denota-se por P_p a variável aleatória referente a posição no timeline que uma mensagem é respondida (reply) e por P_t a mesma variável para o caso de retweets.

Para iniciar a análise, a Figura 4.3a apresenta as CCDFs de P_p e P_t para S_1. Esta amostra foi escolhida pois ela representa com mais fidelidade o todo do conjunto de usuários coletados. Com esta figura percebe-se que em muitos casos usuários buscam mensagens longe do topo do timeline para respondê-las ou compartilhá-las. Por exemplo, 10% de todos os retweets são feitos para tweets que estão em posições superiores a 800 do timeline. Além disso, esta figura permite ver que os usuários não buscam por mensagens tão longe do topo para responder assim como o fazem para compartilhar.

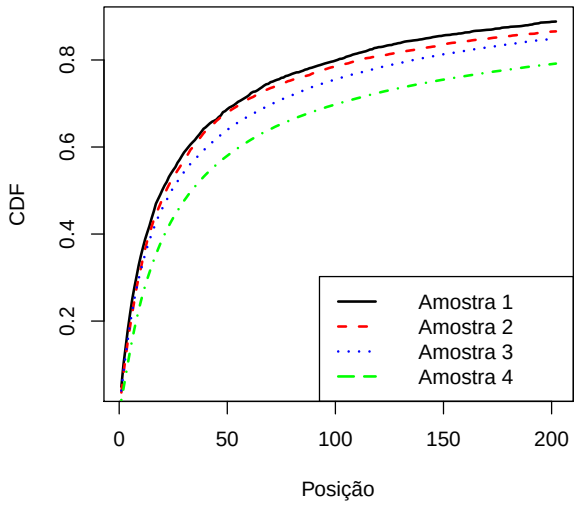
As Figuras 4.3b e 4.3c apresentam a Função de Distribuição Acumulada (CDF)



(a) Distribuições de P_p e P_t para amostra S_1



(b) Distribuições de P_p



(c) Distribuições de P_t

Figura 4.3: Análise das distribuições de P_p e P_t.

de P_p e P_t para todas as amostras consideradas, no entanto, apenas valores menores ou iguais a 200 foram considerados para possibilitar melhor visualização. Pode-se perceber que mensagens mais novas tem mais chance de serem respondidas e compartilhadas (assim como discutido no parágrafo anterior). Por exemplo, 14% dos replies ocorrem na posição 0 para S_1. Essa fração se reduz para 12%, 10% e 5% para S_2, S_3 e S_4 respectivamente. Além disso, 84% (81%, 80%, 77%) de todos os replies acontecem no top-50 para S_1 (S_2, S_3, S_4).

O mesmo fenômeno é observado para retweets, onde tem-se que 68% (68%, 64%,

58%) de todos os retweets estão no top 50 para S_1 (S_2, S_3, S_4). Além disso, as Figuras4.3b e 4.3c mostram que usuários mais ativos (aqueles com mais tweets) tem uma maior chance de responder ou compartilhar mensagens longe do topo do timeline, indicando que usuários que enviam mais tweets também passam mais tempo lendo e interagindo as mensagens recebidas.

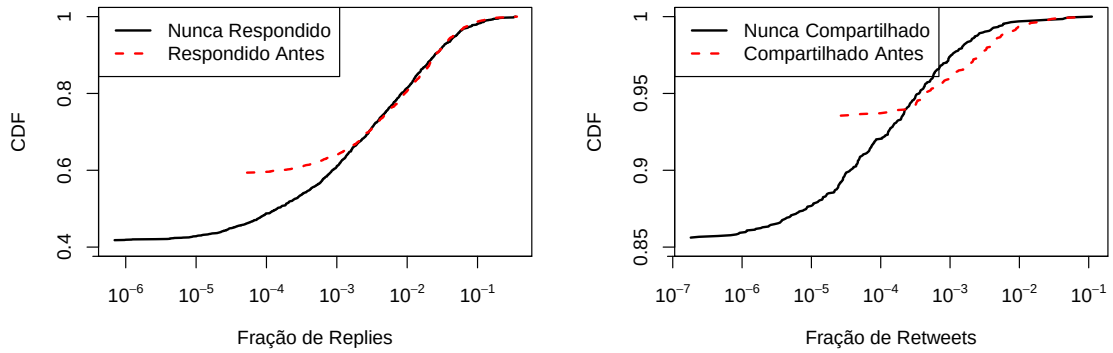
4.3.2 Interações passadas

O objetivo desta seção é verificar a influência de interações passadas entre usuários na ocorrência de interações futuras. Em outras palavras, o objetivo é responder a seguinte questão: *usuários que tiveram mensagens previamente respondidas (compartilhadas) tem uma maior chance de terem mensagens respondidas (compartilhadas) novamente?*

Para responder a essa questão o seguinte processo de medição foi realizado: para cada usuário i na amostra S_k ($k = 1, 2, 3, 4$) e para cada tweet $m \in TL_i$ computou-se a probabilidade condicional de que m seja respondida (compartilhada) por i dado que o usuário que postou m teve uma mensagem respondida (compartilhada) por i anteriormente. A mesma probabilidade foi calculada dado o evento complementar, ou seja, dado que o usuário que postou m não teve uma mensagem respondida (compartilhada) por i anteriormente.

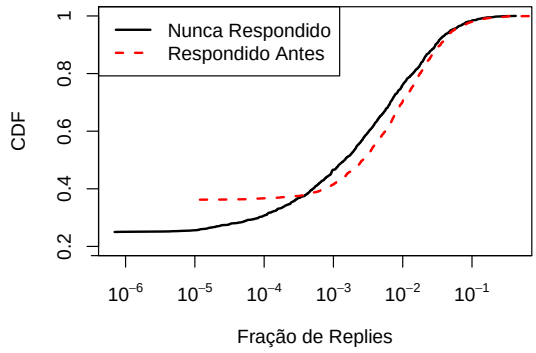
Exemplificando, para calcular a probabilidade de que i responda um usuário dado que i já o respondeu antes, foi contato o número de ocorrência de dois eventos relacionados ao usuário i: A, o número de vezes que i respondeu uma mensagem que veio de um usuário previamente respondido, e B o número de mensagens recebidas que vieram de usuários que i respondeu previamente. Após isso, calcula-se a fração $\frac{A}{B}$.

A Figura 4.4 apresenta um conjunto resultados relacionados com as 4 amostras e com os dois tipos de interações consideradas, replies e retweets. O eixo X de cada figura dá a fração de replies (retweets) para cada usuário da amostra, representando as probabilidades descritas nos parágrafos anteriores. O eixo Y por sua vez, representa a Função de Distribuição Acumulada (CDF) da fração apresentada no eixo X. A primeira coluna da figura é composta por resultados referentes a replies, enquanto a segundo é referente a retweets. Além disso, cada figura possui duas curvas: a primeira em linha sólida com legenda "Respondido antes" (ou "Compartilhado Antes") é referente a probabilidade de se responder (compartilhar) um tweet de um usuário que já teve um tweet respondido (compartilhado) antes, e será denotada por p_p (p_t). A segunda, com legenda "Nunca Respondido" ("Nunca Compartilhado") é referente a probabilidade de se responder (compartilhar) um tweet de um usuário que nunca teve tweet respondido (compartilhado) antes, e será denotada por $\bar{p}_p$ ($\bar{p}_t$).

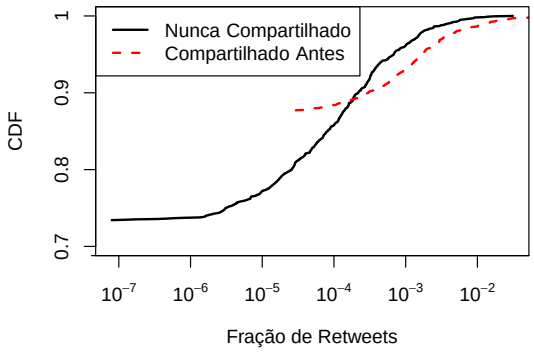


(a) Amostra S_1, Reply

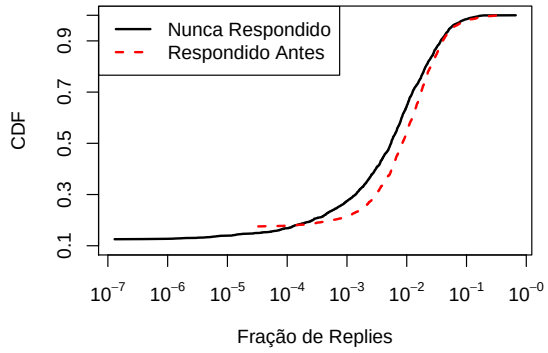
(b) Amostra S_1, Retweet



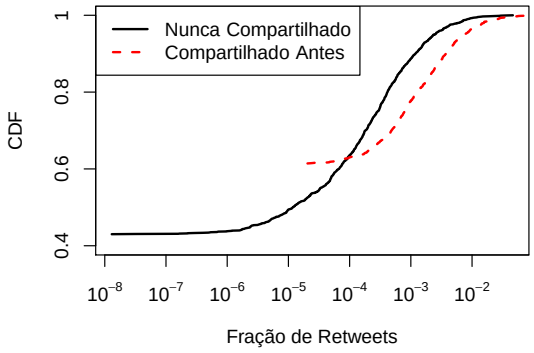
(c) Amostra S_2, Reply



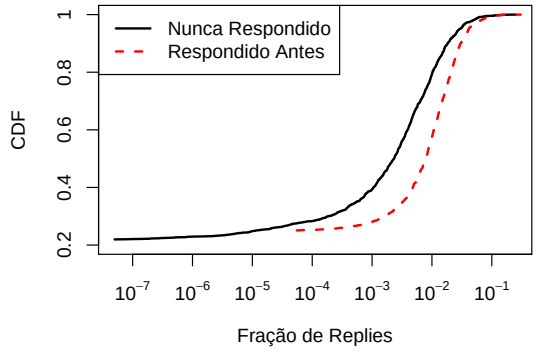
(d) Amostra S_2, Retweet



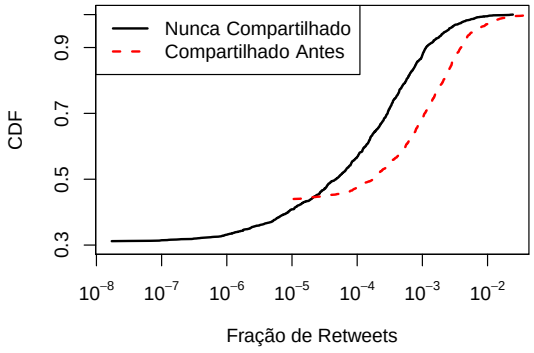
(e) Amostra S_3, Reply



(f) Amostra S_3, Retweet



(g) Amostra S_4, Reply



(h) Amostra S_4, Retweet

Figura 4.4: Probabilidade de interações futuras condicionada a existência de interações passadas.

Os resultados contidos nesta figura podem ser analisado em dois aspectos: o primeiro é o ponto no eixo Y onde cada curva inicia, o qual pode ser facilmente explicado por uma questão de escala, dado que o valor zero não pode ser representado na escala logarítmica. Por exemplo, pela Figura 4.4a (amostra S_1), a probabilidade de um usuário responder outro que nunca respondeu antes é aproximadamente 0.4, ou seja, cerca de 40% dos usuários dessa amostra nunca responderam ninguém. Analogamente, pela mesma figura percebe-se que 60% dos usuários nunca responderam um usuário que já haviam respondido previamente.

Pelo conjunto de figuras também é possível ver que a distribuição de probabilidade representada por p_p (p_t) é deslocada para a direita em relação a $\bar{p}_p$ ($\bar{p}_t$), indicando que as taxas de replies (retweets) ocorrem com maior frequência para usuários que previamente tiveram tweets respondidos (compartilhados). Ademais, a medida que o número de tweets mínimo de cada amostra aumenta, pode-se ver que a diferença entre as duas curvas de cada figura se torna mais evidente.

Pelo exposto, pode-se concluir que condicionado ao fato de que um usuário i respondeu (compartilhou) uma mensagem de j no passado, então existe uma maior chance de que i interaja com j novamente do que com um usuário que nunca tenha o feito.

Estes resultados sugerem uma relação do comportamento dos usuários do Twitter com os estudos de Dunbar [1993]. Segundo o autor, existe um limite cognitivo para o número de pessoas que um indivíduo pode manter relações sociais estáveis. No contexto do Twitter isto implica que mesmo seguindo um grande número de pessoas não é comum interagir frequentemente com todas elas, mas sim com um subgrupo. Como consequência disso existe uma maior chance de se responder (compartilhar) usuários que tiveram mensagens previamente respondidas (compartilhadas) assim como caracterizado nesta seção.

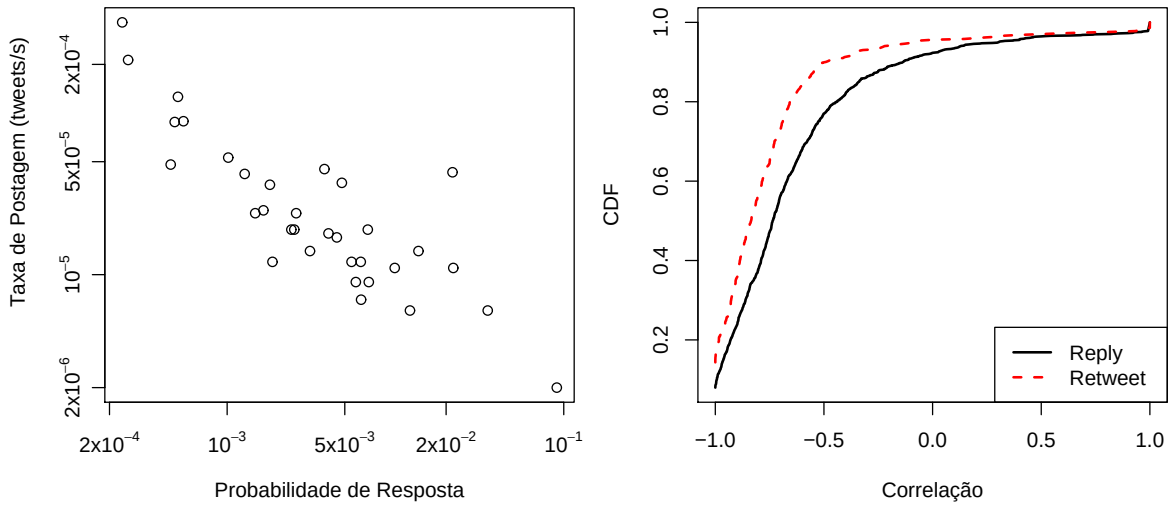
4.3.3 Taxa de postagem

A próxima característica a ser analisada é o nível de atividade do usuário que está enviando tweets. Mais especificamente, tem-se o interesse em responder a seguinte questão: *Usuários mais ativos tem mais chances de terem tweets respondidos (compartilhados)?*

Para responder essa pergunta, procedeu-se da seguinte forma: para cada usuário $i \in S_k$ ($k = 1, 2, 3, 4$) e para cada $j \in Out_i$ comparou-se a taxa de envio de tweets de j com a fração de suas mensagens que foram respondidas (compartilhadas) por i. É importante ressaltar que a taxa de envio de tweets de j é definida pela razão entre o número de tweets que postou pelo intervalo entre seu primeiro e último tweet.

A Figura 4.5a mostra a relação entre essas duas variáveis para um usuário tomado com exemplo de i. Cada ponto da figura corresponde a um usuário $j \in Out_i$. Usuários que nunca foram respondidos por i foram omitidos devido a escala logarítmica. A figura indica que quanto maior a taxa de envio de tweets de j, menor é a probabilidade de que i o responda. Com intuito de verificar se esta relação é comum entre os usuários do Twitter este procedimento foi repedido para todo $i \in S_k$ para então calcular-se a correlação linear entre as variáveis taxa de envio dos usuários $j \in Out_i$ e suas respectivas probabilidades de resposta por parte de i (em escala logarítmica).

A Figura 4.5b apresenta o resultado destes cálculos para a amostra S_2. As demais amostras apresentam o mesmo comportamento e por isso suas respectivas figuras foram omitidas. Pode-se perceber que no caso de replies (retweets) quase 80% (90%) de todos os usuários originam um coeficiente de correlação menor do que -0.5, ao passo que apenas 10% (9%) geram uma correlação positiva. Estes resultados indicam que em geral os usuários do Twitter tem menos interesse em responder pessoas com taxas de envio de tweets elevadas, assim como exemplificado na Figura 4.5a.



(a) Caso de exemplo

(b) Correlação entre log(taxa de envio de tweets) e log(probabilidade de resposta)

Figura 4.5: Análise da influência da taxa de envio de tweets na probabilidade de resposta de um usuário.

4.3.4 Atributos textuais

Nesta seção é investigada a influência de características intrínsecas de mensagens nos padrões de interações, representadas por replies ou retweets, entre usuários. O objetivo

não é olhar para a semântica dos tweets, mas sim, conduzir esta análise considerando o comprimento de cada tweet (número de caracteres) a presença de *hashtags*, *mentions* e *links* de páginas *Web*, (i.e. URLs). Para essa finalidade todos os tweets da coleção de dados foram analisados com relação a essas características após serem separados em três categorias: *i)* tweets que foram respondidos; *ii)* tweets que foram compartilhados; e *iii)* todos os tweets.

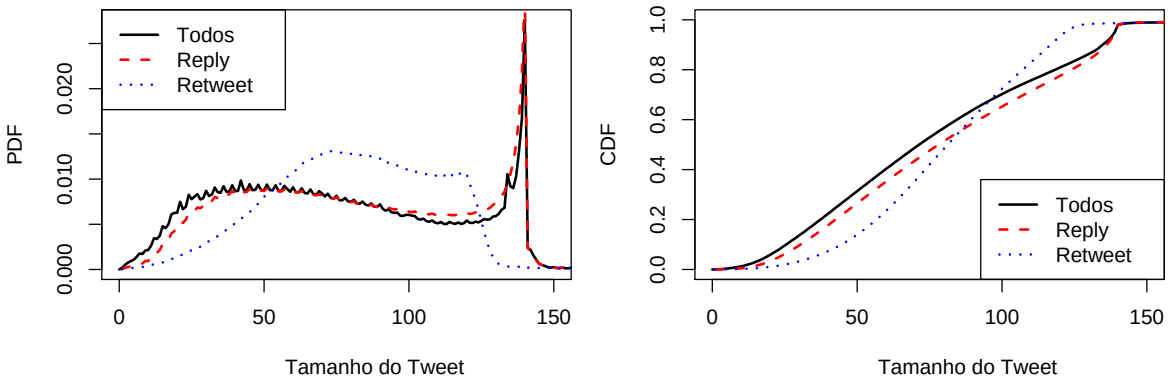
A Tabela 4.2 apresenta a fração de tweets com *hashtags*, *mentions* ou URLs nestas três categorias da coleção de dados. Por um lado, pode-se perceber que tweets com *hashtags* e *mentions* tem uma chance maior de serem compartilhados. Por outro, observa-se que este comportamento não sem mantém para tweets com *mentions*, uma vez que apenas 25% das mensagens compartilhadas contém uma menção a outro usuário. Para o conjunto de mensagens respondidas ocorre exatamente o contrário, ou seja, a maior parte destes (cerca de 55%) contém *mentions*, enquanto apenas uma pequena fracão contém *hashtags* ou URLs.

Tabela 4.2: Fração de tweets com *hashtags*, *mentions* e URLs

	Respondidos	Compartilhados	Todos
Hashtag	0.04	**0.16**	0.05
Mention	**0.55**	0.25	0.36
URL	0.10	**0.51**	0.22

A próxima característica analisada foi o comprimento do tweet (em número de caracteres). As Figuras 4.6a e 4.6b apresentam a Função de Distribuição de Probabilidade (PDF, do inglês *Probability Distribution Function*) e Função de Distribuição Acumulada (CDF) do comprimento dos tweets para os três conjuntos analisados. A primeira observação interessante é que foram encontrados na base algumas mensagens com mais de 140 caracteres. Foram apenas poucos casos, e uma possível explicação seria alguma falha do sistema no início do Twitter.

Examinando essas figuras pode-se notar que a PDF e a CDF do conjunto de tweets respondidos é muito similar ao conjunto de todas as mensagens da coleção de dados. No entanto, a distribuição para o conjunto de mensagens compartilhadas tem um comportamento completamente diferente dos demais conjuntos. Neste caso, a Figura 4.6a não apresenta nenhum pico próximo de 140 caracteres. Ao invés disso, percebe-se a forma aproximada de uma distribuição uniforme no intervalo de 50 a 120 caracteres. Após este intervalos, percebe-se uma queda na probabilidade de compartilhamento. Este último fato pode ser explicado pelo fato de que na época que os dados foram coletados era necessário que a mensagem possuísse espaço suficiente para inserir o



(a) Distribuição de Probabilidade (PDF) (b) Distribuição Acumulada (CDF)

Figura 4.6: Caracterização da importância do número de caracteres do tweet na probabilidade de interação.

texto característico de um retweet (isto é, `RT @apelido`).

4.4 Discussões

Este capítulo apresentou uma descrição do comportamento dos usuários do Twitter no que tange a intensidade com que estes postam mensagens na rede e fatores que influenciam interações entre usuários. Em geral eles preferem responder e compartilhar mensagens que não sejam antigas, que venham de usuários com taxas e envio de mensagens não elevada e com quem já tenham interagido antes. Tem-se também que mensagens com menções a usuários e com comprimento próximo a 140 caracteres têm maiores chances de serem respondidas, enquanto as que possuem URLs, *hashtags* e comprimento inferior a 120 caracteres são mais prováveis de serem compartilhadas. Não obstante, foram apresentadas evidências de que os usuários podem estar recebendo mais informações do que são capazes de lidar. Por exemplo, muitas vezes eles buscam por centenas de mensagens até encontrarem uma que tem interesse em responder ou compartilhar. Todos estes resultados claramente motivam o projeto de ferramentas que possam facilitar o uso do Twitter por parte dos usuários, tema este, que será abordado no próximo capítulo.

Capítulo 5

Uma Metodologia para Reorganização do Timeline

O objetivo deste capítulo é apresentar e avaliar uma metodologia alternativa para reorganização do timeline do Twitter. A Seção 5.1 apresenta a metodologia em detalhes, a qual, é composta basicamente de uma heurística para detectar quando usuários estão ativos (postando mensagens) e do uso de dois algoritmos de aprendizado de máquina. Após isso, a Seção 5.2 mostra o procedimento experimental usado para avaliar a metodologia proposta e discute os resultados encontrados. Por fim, a Seção apresenta uma discussão do conteúdo deste capítulo.

5.1 Estratégia de Reorganização

O Capítulo 4 mostrou que os usuários do Twitter podem passar um longo tempo lendo tweets em seus timelines até que encontrem um que tenham interesse de responder ou compartilhar. Além disso, foram apresentadas características que indicam quais tipos de tweets são mais interessantes para os usuários. Motivado por estes resultados, nesta seção será apresentada uma estratégia que objetiva reorganizar o timeline usual do Twitter. A meta desta reorganização é apresentar os tweets com maiores chances de serem respondidos ou compartilhados (e portanto aqueles com maior potencial de serem interessantes) no topo do timeline.

Embora sete características relacionadas com taxas de interações no Twitter tenham sido exploradas no capítulo anterior, apenas três serão utilizadas na metodologia aqui apresentada, a saber: a idade do tweet (medida através da distância que o tweet está do topo do timeline), a taxa de postagem de tweets do usuário que postou um determinado tweet em análise e interações passadas, também considerando o usuário

que originou o tweet que está sendo analisado. Embora pareça um contrassenso ter investigado características que não serão utilizadas, as referentes aos atributos textuais, existiram motivos para que esta decisão fosse tomada, entre eles: como visto na Seção 4.3.4, a presença de *hashtags*, *mentions*, URLs e o tamanho da mensagem não são igualmente importantes para taxas de replies e retweets, o que prejudica o desenvolvimento de uma metodologia genérica o suficiente para tratar este dois tipos de interação ao mesmo tempo. Além disso, segundo um estudo qualitativo realizado por André et al. [2012] o uso indiscriminado de alguns elementos textuais em tweets, como os aqui apresentados, pode fazer com estes tweets se tornem "chatos" na opinião dos usuários.

As duas abordagens a serem descritas nesta seção são instâncias do procedimento geral descrito pelo Algoritmo 1. A ideia principal é reconhecer que os usuários podem estar em dois diferentes estados no que se diz respeito as suas interações com seus respectivos timelines do Twitter: *online* (ou ON), quando estão prestando atenção em seus timelines; e *offline* (ou OFF), em caso contrário. Considera-se que quando um usuário está no estado ON, ele está vendo todos os tweets que recebeu durante sua última sessão OFF e também os tweets que está recebendo durante a corrente sessão ON. Baseado neste comportamento típico dos usuários a proposta consiste em realizar o processo de reorganização do timeline na ocorrência de dois eventos: *i)* em toda mudança de estado de OFF para ON (Linha 2 do algoritmo); e *ii)* toda chegada de tweet no timeline (Linha 5). Após cada processo de reorganização, o timeline modificado (TL') deve ser apresentado ao usuário ao invés do antigo (TL).

Algoritmo 1: Procedimento Geral de Reorganização do Timeline

Data: Usuário u e Tweets que u recebeu em sua última sessão OFF
1 $TL \leftarrow$ Conjunto de todos Tweets que u recebeu em sua última sessão OFF
2 $TL' \leftarrow$ Reorganize(TL)
3 **foreach** *Tweet m recebido na sessão ON corrente* **do**
4 $\quad TL \leftarrow TL \cup \{m\}$
5 $\quad TL' \leftarrow$ Reorganize(TL)
6 **end**

As instâncias deste algoritmo são obtidas com duas versões diferentes do procedimento Reorganize(). Ambas são baseadas em algoritmos de aprendizado de máquina, as quais são utilizadas para calcular pontuações para os tweets do timeline tendo como base as três características referidas anteriormente de cada um destes tweets. Estas técnicas são os classificadores *Naive Bayes* (NB) e *Support Vector Machine* (SVM) os quais serão apresentados nas Seções 5.1.2 e 5.1.3 respectivamente. Após

computar pontuações para os tweets do timeline, estes são ordenados de forma que os que possuem maiores chances de serem respondidos ou compartilhados (os com maior pontuação) sejam apresentados para o usuário no topo do timeline.

Um passo importante da metodologia é que ela considera que as pontuações de cada tweet podem variar com o tempo uma vez que o procedimento Reorganize() pode ser realizado várias vezes durante um período de atividade do usuário (um sessão ON). Desta forma, um tweet que é interessante em um dado momento pode não o ser mais no futuro. Obviamente, essa propriedade do algoritmo vem com um custo computacional agregado. Em termos práticos, pode ser muito "caro" reorganizar um timeline por completo toda vez que um novo tweet chegar. Para tentar contornar este problema, o segundo tipo de evento (chegada de um tweet) pode ser substituído por algum mecanismo de *time out*, ou simplesmente pela chegada de $k > 0$ tweets no timeline.

Uma vez que o conjunto de dados analisado neste trabalho não permite saber com exatidão quando os usuários estão vivenciando uma sessão ON ou OFF, um simples modelo foi usado para tentar inferir este comportamento. Este modelo é apresentado na próxima seção.

5.1.1 Identificação de Períodos de Atividade

Nesta seção é apresentado um simples modelo com intuito de descrever padrões de atividades entre o Twitter e seus usuários. Como descrito na seção anterior, cada usuário pode estar em dois estados: ON ou OFF. De acordo com este modelo será assumido que um usuário está no estado ON durante o intervalo de tempo que estiver ativamente postando mensagens no Twitter, de forma que o tempo entre dois destes eventos consecutivos não exceda um determinado limiar T_{OFF}. No restante do tempo, define-se que o usuário está no estado OFF. A Figura 5.1 ilustra a transição entre esses dois estados ao logo do tempo para um usuário de acordo com o modelo adotado. É importante observar que esta definição tem a limitação de não considerar as atividades passivas dos usuários, tais como a simples leitura de tweets.

Uma vez que os dados descritos no Capítulo 3 não contém explicitamente os delimitadores dos estados ON e OFF, o número de sessões depende do valor do parâmetro T_{OFF} adotado. Seguindo o procedimento apresentado por Menascé et al. [1999], o valor deste parâmetro foi variado, e em cada caso foi contado o número de sessões originado. Um valor muito pequeno (por exemplo, um minuto) poderia resultar em um volume muito grande de sessões. A medida que se aumenta o valor de T_{OFF}, o número de sessões é reduzido continuamente até que se estabilize. Este valor de T_{OFF}

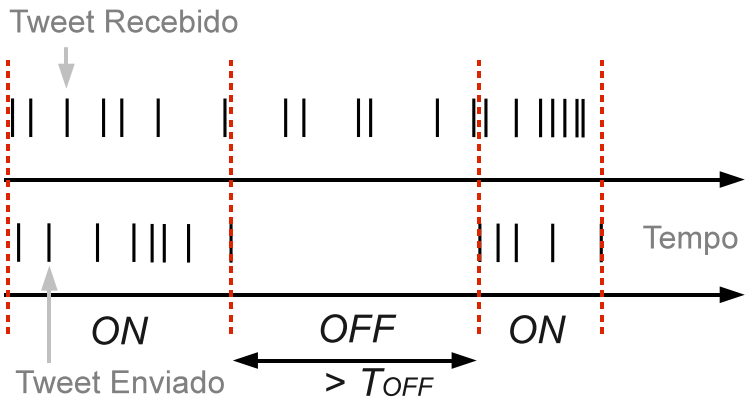
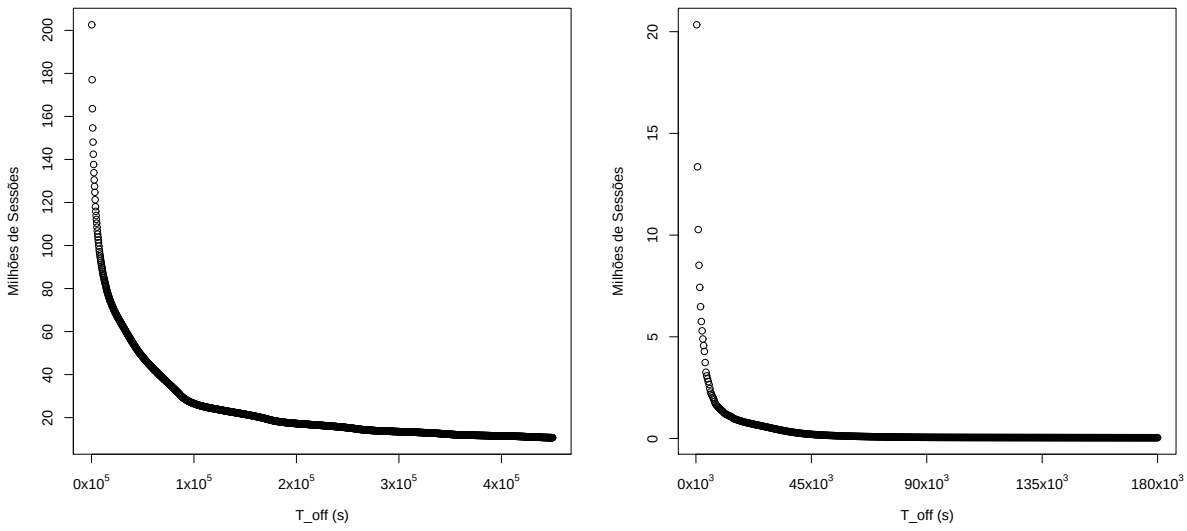


Figura 5.1: Ilustração dos estados *ON* e *OFF*.

é então adotado para o modelo. Durante os experimentos conduzidos para a identificação deste parâmetro apenas os usuários com mais de mil mensagens postadas foram considerados. Esta decisão foi tomada pois ao considerar usuários com poucos tweets a variância do número total de sessões era grande a ponto de levar a estimativas muito grandes e provavelmente imprecisas de T_{OFF}, assim como ilustrado na Figura 5.2a. O resultado final desta medição mostrou que o número de sessões se estabiliza quando T_{OFF} está próximo de dez mil segundos, o que representa aproximadamente três horas (Figura 5.2b).



(a) Usuários com mais de 2 tweets

(b) Usuários com mais de 1000 tweets

Figura 5.2: Caracterização do número de sessões em função de T_{OFF}.

Comparado com outros trabalhos que caracterizaram sessões de interações em

sistemas *Web* tradicionais [Arlitt, 2000; Oke & Bunt, 2002], os valores de T_{OFF} encontrados são maiores do que os típicos 10-45 minutos usualmente observados. A razão mais intuitiva para para este comportamento é o longo período de tempo que os usuários do Twitter passam utilizando a rede e tentando se atualizar de todos os eventos que ocorrem em tempo real.

É importante ressaltar que o modelo ON-OFF descrito consiste apenas de uma heurística para inferir quando usuários estão ON ou OFF. Esta estratégia foi adotada devido a limitações da coleção de dados. Em uma situação real, se existir uma informação mais confiável sobre o estado do usuário, esta deverá ser utilizada.

5.1.2 Abordagem usando *Naive Bayes*

Esta seção apresenta como o classificador *Naive Bayes* [Hastie et al., 2009], denotado aqui por NB, foi utilizado para calcular pontuações para tweets no procedimento REORGANIZE() do Algoritmo 1. Neste caso, a pontuação é representada através de medida de probabilidade associada a cada tweet. Medida esta que representa a probabilidade de se interagir (responder ou compartilhar) com um tweet dado seu conjunto de atributos. Como previamente dito, foram considerados três atributos:

- $Age(m)$, a idade do tweet m medida através de sua posição no timeline (para o mais novo $Age(m) = 0$, para o próximo 1, e assim por diante);
- $SR(m)$, a taxa média de envio de tweets do usuário que enviou m; e
- $I(m)$, um indicador binário que pode ser 1, se o usuário já interagiu com quem enviou m antes e 0 em caso contrário.

Desta forma, para cada tweet m, sua pontuação é definida por:

$$P(m) = P(\text{Interagir com } m | Age(m) = p, SR(m) = r, I(m) = b). \tag{5.1}$$

Sob a hipótese ingênua, inerente ao classificador *Naive Bayes*, de que os eventos $Age(m) = p$, $SR(m) = r$, $I(m) = b$ são independentes entre si, tem-se:

$$\begin{aligned} P(m) &= P(\text{Interagir com } m |\ Age(m) = p) \\ &\times\ P(\text{Interagir com } m |\ SR(m) = r) \\ &\times\ P(\text{Interagir com } m |\ I(m) = b). \end{aligned}$$

O desafio passa a ser então como calcular cada um dos fatores deste produto, ou seja, como calcular individualmente a pontuação de cada tweet para cada característica considerada. As próximas seções apresentam os modelos considerados para cada caso. Além disso, utilizando uma amostra aleatória S de U com dois mil usuários mostra-se, quando aplicável, porque os modelos adotados são adequados..

5.1.2.1 Pontuação associada com a Idade

Suponha que um tweet m está na posição p do timeline de uma dado usuário. Então, assume-se o seguinte modelo:

$$P(\text{Interagir com } m|\ Age(m) = p) = \begin{cases} \beta_1 p^{\alpha_1}, & p \leq 10 \\ \beta_2 p^{\alpha_2}, & p > 10. \end{cases} \tag{5.2}$$

Para explicar a escolha deste modelo, a Figura 5.3 apresenta a probabilidade estimada de ocorrer um reply, retweet ou um dos dois como uma função de p para a amostra S. Pode-se perceber que o modelo proposto tem um bom ajuste com os dados empíricos nas três figuras e para os dois regimes identificados, $p \leq 10$ e $p > 10$, justificando assim que é uma escolha simples e adequada.

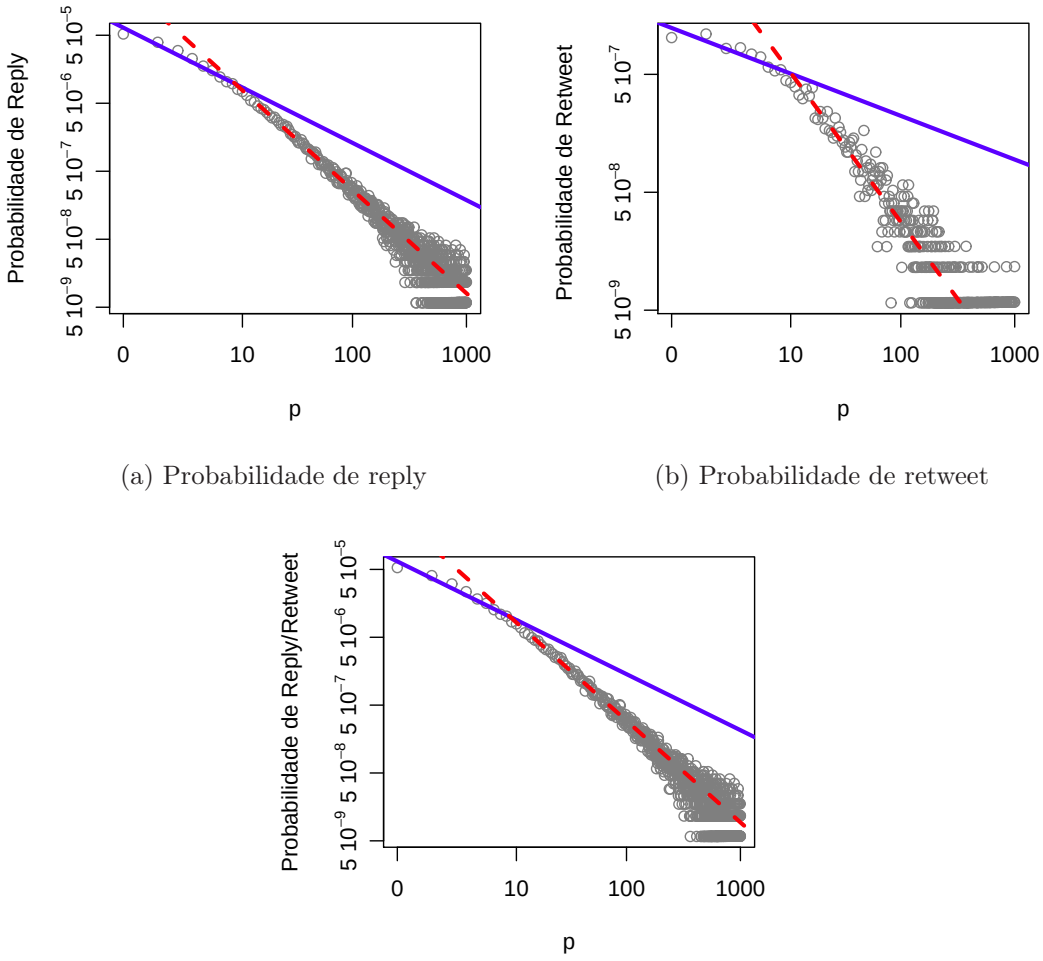
Ressaltando que para calcular as probabilidades apresentadas nas referidas figuras, o seguinte procedimento foi realizado: assuma que o tweet m está na posição p do timeline do usuário u. Então, u teve a chance de interagir com m quando ele esteve nas posições $0, \ldots, p-1$ e tem no referido instante a chance de interagir com ele na posição p. Desta forma, computou-se a fração de tweets respondidos (compartilhados) na posição p de todos aqueles tweets que poderiam ter sido respondidos (compartilhados) nesta posição.

5.1.2.2 Pontuação associada com a Taxa de Envio

Suponha que um tweet m tenha sido enviado por um usuário que envia tweets a uma taxa r. Então assume-se o seguinte modelo:

$$P(\text{Interagir com } m|\ SR(m) = r) = \beta r^{\alpha}. \tag{5.3}$$

Para justificar a escolha deste modelo procedeu-se de forma análoga a seção anterior. A Figura 5.4 apresenta a probabilidade de reply (retweet) em função de r e as respectivas curvas ajustadas para cada caso. Pode-se perceber que o modelo proposto tem um bom ajuste com os dados observados.



(a) Probabilidade de reply

(b) Probabilidade de retweet

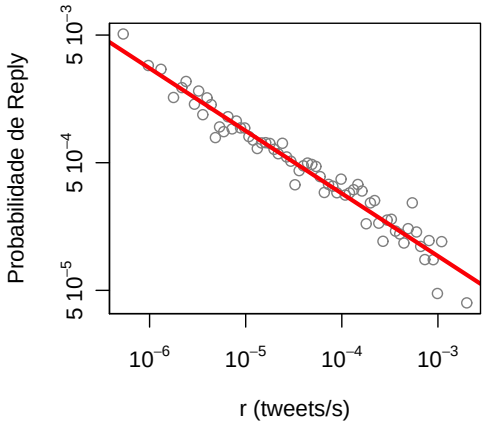
(c) Probabilidade de reply ou retweet

Figura 5.3: Probabilidade de interagir com um tweet dado que sua posição no timeline é p. Escala logarítmica em ambos os eixos.

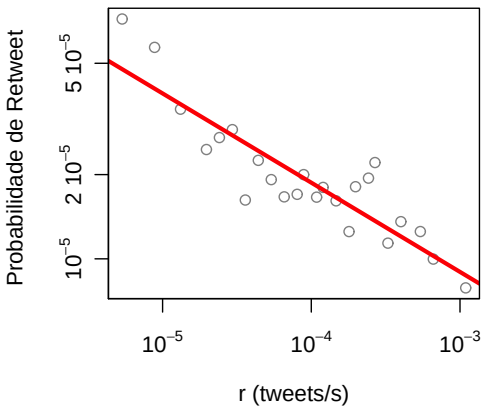
Para calcular as probabilidades apresentadas nas referidas figuras o seguinte procedimento foi realizado: Pada todos os usuários $u \in S$ calculou-se a fração de replies (retweets) feitos para usuários com uma taxa de envio de tweets r, de todos os tweets que foram recebidos no timeline originados de usuários com esta mesma taxa de envio. Uma vez que $SR(m)$ é uma variável contínua, os valores computados de r foram agrupados em *bins* logarítmicos e estes *bins* foram agregados de forma que cada um tivesse ao menos 100 observações.

5.1.2.3 Pontuação associada com Interações Passadas

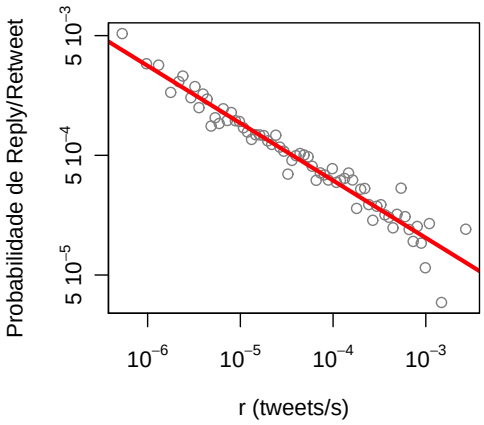
Uma vez que $I(m)$ é uma variável binária, decidiu-se utilizar o seguinte modelo:



(a) Probabilidade de Reply



(b) Probabilidade de Retweet



(c) Probabilidade de Reply ou Retweet

Figura 5.4: Probabilidade de interagir com um tweet dado que a taxa de envio de que o originou é r. Escala logarítmica em ambos os eixos.

$$P(\text{Interagir com } m|\ I(m) = b) = \begin{cases} \gamma_1, & b = 1 \\ \gamma_2, & b = 0. \end{cases} \tag{5.4}$$

Para computar os valores de γ_1 e γ_2 a seguinte medição foi realizada: para todos usuários $u \in S$, para cada chegada de tweet em seu timeline computou-se a fração de tweets m que foram respondidos (compartilhados) por u tais que $I(m) = b$ de todos os tweets que foram recebidos com $I(m) = b$.

5.1.3 Abordagem usando *Support Vector Machine*

Support Vector Machine (SVM) é um conjunto de métodos muito úteis e vastamente utilizados em problemas de classificação de dados. A versão mais comum de SVM é um

classificador supervisionado binário que mapeia um vetor de atributos em duas classes. Nesta dissertação esta versão de SVM foi utilizada, mais especificamente a apresentada por Hsu & Lin [1999], para a implementação do procedimento Reorganize() no Algoritmo 1. A ideia é composta de dois passos: primeiro, usar os mesmos atributos considerados na seção anterior para classificar os tweets do timeline em *interessantes* (mais prováveis de se interagir com) e não interessantes. Em outras palavras, para cada tweet m do timeline do usuário o classificador SVM foi utilizado para mapear o vetor $[Age(m), SR(m), I(m)]$ em 1 (interessante) ou 0 (não interessante).

O segundo passo consiste de apresentar primeiro ao usuário os tweets classificados como 1 em ordem cronológica reversa e depois os classificados como 0, também em ordem cronológica reversa. É importante ressaltar que uma vez que os atributos dos tweets podem mudar com o passar do tempo, o mesmo pode acontecer com a classe que pertencem.

5.2 Resultados Experimentais

Nesta seção são apresentados resultados referentes a avaliação dos algoritmos propostos. Primeiro, a metodologia de avaliação é apresentada, para então discutir os resultados. Após isso, é avaliado o impacto do parâmetro T_{OFF}, introduzido na Seção 5.1.1, e finalmente, os resultados são discutidos quando os algoritmos são aplicados em conjuntos de usuários com padrões de atividades completamente diferentes.

5.2.1 Metodologia Experimental

A metodologia escolhida para avaliar os algoritmos propostos consiste de um processo de simulação orientada a dados. A ideia é entender o que iria ocorrer se os tweets tivessem sido apresentados para os usuários de acordo com os algoritmos propostos nesta dissertação ao invés da ordem cronológica reversa, comumente utilizada pelo Twitter. Os passos deste processo de simulação são mostrados no Algoritmo 2 e discutidos a seguir.

Para um determinado usuário u, seus tweets são divididos em listas de sessões, uma para sessões ON e outra ara OFF, de acordo com a Seção 5.1.1 (Linha 1). Após isso, para cada tweet que for uma interação (uma interação pode ser um reply ou retweet) de u, busca-se pela origem deste tweet (aquele que foi respondido ou compartilhado) no timeline de u. No entanto, busca-se apenas entre os mais recentes, ou seja, aqueles que foram recebidos desde o início da última sessão OFF até o tempo em que a interação foi realizada (Linhas 5 e 8). Antes da realização da busca (Linha 10) o

Algoritmo 2: Procedimento de Simulação

Data: Usuário u, M_u, TL_u, Out_u

Divida M_u em duas listas, uma para sessões ON (on_u) e outra para as sessões OFF (off_u)
foreach *session* $s \in on_u$ **do**
    $t_1 \leftarrow$ instante de tempo em que s inicia
    $t_2 \leftarrow$ instante de tempo em que se inicia a última sessão OFF em off_u que precede s
    $TL \leftarrow$ lista de tweets em TL_u que foram postados depois de t_1 e antes de t_2
    $IT \leftarrow$ lista de replies e retweets em s
    foreach $m \in IT$ **do**
        Atualize TL com as mensagens postadas antes de m
        $TL' \leftarrow$ Reorganize(TL)
        Procure pelo tweet ao qual m responde (compartilha) em TL'
        Retorne a posição em que estava a origem de m em TL', quando possível
    end
end

procedimento de reorganização é feito assim como discutido nas Seções 5.1.2 ou 5.1.3. Se o tweet respondido (compartilhado) for encontrado no timeline de u a posição em que ele estava é retornada. É importante ressaltar que não necessariamente o tweet será encontrado no timeline, pois é possível que ele seja um reply (retweet) realizado para um usuário que u não segue.

Repetindo este procedimento para uma grande quantidade de usuários da coleção de dados é possível comparar a metodologia proposta nesta dissertação com o timeline usual do Twitter. Para isso, decidiu-se comparar a fração de tweets respondidos (compartilhados) nas p primeiras posições do timeline quando o método de organização do timeline é a ordem cronológica reversa ou um dos propostos.

Com o intuito de ter uma boa estimativa destes valores é importante executar esta simulação para uma grande quantidade de usuários da coleção de dados. No entanto, devido ao alto custo computacional, este procedimento não pôde ser realizado para todo o conjunto U. Para contornar este problema, decidiu-se trabalhar com amostras aleatórias. Para isso, foi extraída uma amostra aleatória de U com 10 mil usuários. Associados a esta amostra tem-se 2.25 milhões de tweets, dois quais 540 mil são replies e 62 mil retweets. Além disso, os usuários desta amostra seguem um total de 500 mil outros usuários os quais postaram um total de 200 milhões de tweets.

Ao trabalhar com amostras aleatórias é importante apresentar garantias de que o número de observações seja o suficiente para prover significância estatística para os resultados. Logo, deve-se garantir que o número de replies e retweets citados no

parágrafo anterior são superiores a um determinado limiar. Utilizando técnicas de amostragem (ver Apêndice B), de acordo com Cochran [1977] uma aproximação de estimativa pessimista neste caso com 95% de confiança e um erro absoluto não maior que 0.005, seria 38,416, mostrando que a amostra descrita é representativa para estimar a fração de replies (retweets) nas p primeiras posições do timeline[1].

Após isso, essa amostra foi dividida em 5 subamostras, cada uma com 2 mil usuários, com o objetivo de realizar um procedimento de validação cruzada. Para isso, uma subamostra foi retida para obtenção de dados para treinar os algoritmos de classificação e as demais para testes. Este procedimento foi executado 5 vezes, sendo que em cada, uma amostras era retida para treinamento e as demais para testes. Na etapa de teste, calculou-se a fração de replies e retweets nas primeiras p posições do timeline, para $p = 1$, $p = 5$ e $p = 10$. Os resultados das diferentes combinações de treinamentos e testes foram combinados para possibilitar o cálculo de médias e intervalos de confiança.

É importante ressaltar que cada classificador utilizado requer uma etapa de treino distinta. Para o algoritmo NB, o treino é definido como sendo estimar os parâmetros nos modelos de probabilidade apresentados na Seção 5.1.2. Para o classificador SVM, o seguinte processo foi executado: criou-se um conjunto de dados, tal que cada um de seus elementos é composto de três atributos e um rótulo. Para obter estes dados, cada subamostra foi analisada de forma que para cada reply (retweet) encontrado anotou-se o dados referentes ao tweet respondido (compartilhado) com um rótulo 1 e selecionou-se outro tweet no timeline do usuário aleatoriamente que não havia sido respondido (compartilhado), com rótulo 0. Para cada subamostra ambos os algoritmos foram treinados separadamente com apenas informações referentes a replies, retweets e em um terceiro caso com a união destes dois conjuntos. Salientando que, no caso do SVM, foi utilizada a biblioteca *Libsvm* [Chang & Lin, 2011] como base para as simulações. Para este caso, foi utilizada uma ferramente contida na biblioteca que permite encontrar os melhores parâmetros dos modelos.

5.2.2 Resultados Gerais

Nesta seção, são apresentados os resultados obtidos com os algoritmos propostos na Seção 5.1. A Figura 5.5 apresenta as frações de replies e retweets relacionadas com as p primeiras posições do timeline. A primeira coluna (Figuras 5.5a, 5.5c e 5.5e) é formada de resultados referentes a replies enquanto a segunda (Figuras 5.5b, 5.5d e 5.5f) é

[1]O mesmo argumento pode ser utilizado para as amostras utilizadas no Capítulo 4 e Seção 5.1.2, mas com 90% de confiança e um erro absoluto menor que 0.01.

formada de resultados referentes a retweets. Em cada caso existem três subfiguras. As primeiras são os resultados obtidos com os classificadores treinados com informações de replies, a segunda, com informação de retweets e a terceira com informação de replies e retweets.

Na Figura 5.5a pode-se ver que ambos, NB e SVM melhoraram significativamente a fração de tweets respondidos nas primeiras posições do timeline. Além disso, tem-se que NB e SVM tem resultados parecidos e que de acordo com os intervalos de confiança são estatisticamente equivalentes. Usar informações de retweets para treinar os classificadores e testar para replies (Figura 5.5c) mostrou-se não ser uma estratégia adequada uma vez que os resultados obtidos não são bons. Quando usadas informações de replies e retweets para treinar os classificadores (Figura 5.5e) resultados semelhantes aos da Figura 5.5a foram obtidos.

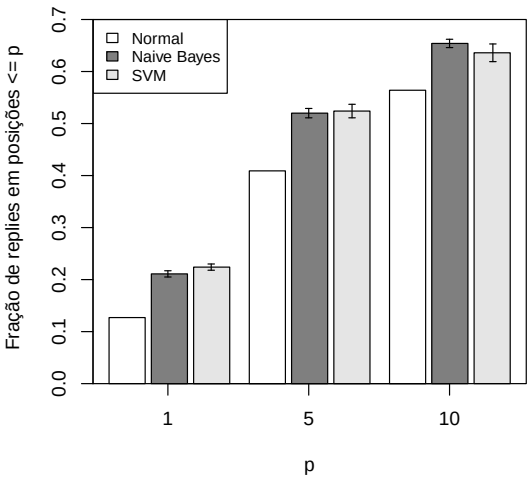
Analisando os resultados relacionados com as frações de retweets, pode-se ver que quando são usadas informações de replies para treinar os algoritmos (Figura 5.5b) a estratégia com NB teve resultados significativos, mas o mesmo não se repetiu para SVM. Na Figura 5.5d é mostrado que NB não origina bons resultados. No entanto, SVM os apresenta, exceto quando $p = 10$. No último caso, (Figura 5.5f) tanto NB quando SVM apresentaram bons resultados os quais demonstram ser estatisticamente equivalentes.

Em geral, pode-se ver que usando informações de replies e retweets para treinar os classificadores bons resultados foram obtidos em termos de apresentação de replies e retweets. Nestes caso, melhorias de aproximadamente 50%, 20% e 10% foram encontrados para p igual a 1, 5 e 10 respectivamente.

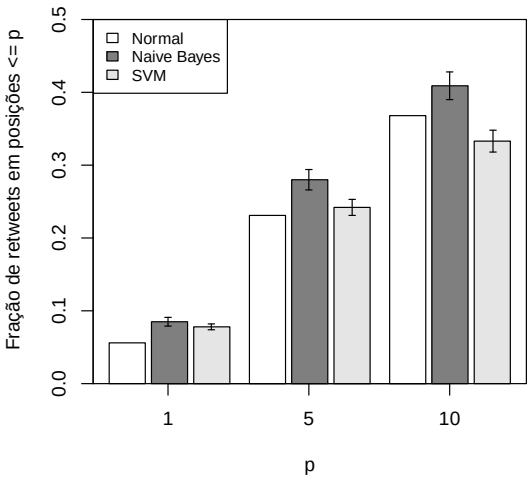
5.2.3 Robustez do Modelo *ON-OFF*

Todos os resultados apresentados na seção anterior usaram o parâmetro T_{OFF} igual a 10 mil segundos no modelo ON-OFF apresentado na Seção 5.1.1. Uma pergunta que naturalmente surge é: *esse é o valor mais indicado para todos os tipos de usuários?* Para responder essa pergunta conduziu-se um experimento no qual T_{OFF} foi variado em 16 valores no intervalo de 10^3s a 10^5s. Neste experimento foram utilizadas duas sub amostras da seção anterior, uma para treinamento e outra para testes. A escolha dessas sub amostras foi feita de forma aleatória. Além disso, é importante ressaltar que informações de replies e retweets foram utilizadas para o treinamentos dos classificadores, uma vez que a seção anterior mostrou evidências de que esta é a melhor estratégia para treinar os algoritmos.

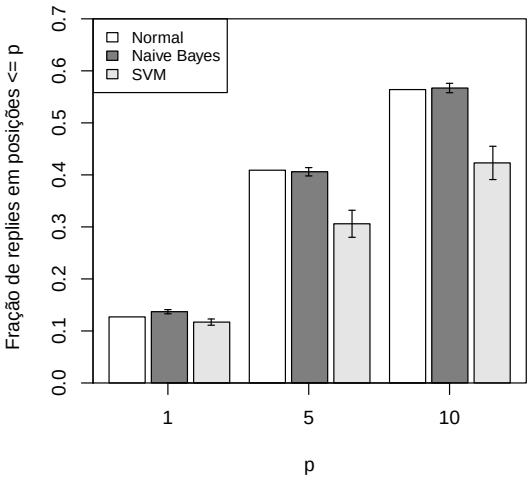
A Tabela 5.1 mostra os resultados deste experimento. Para propiciar uma melhor



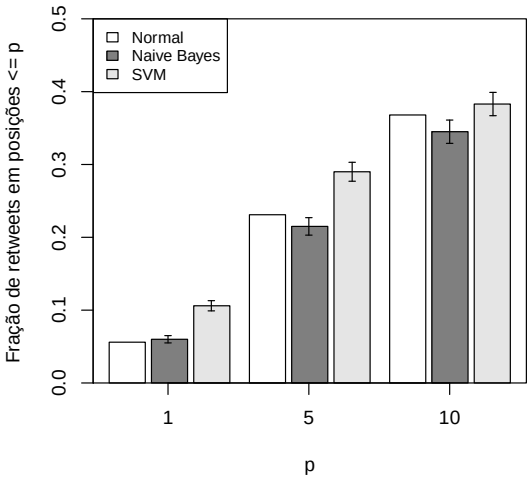
(a) Informação de Replies



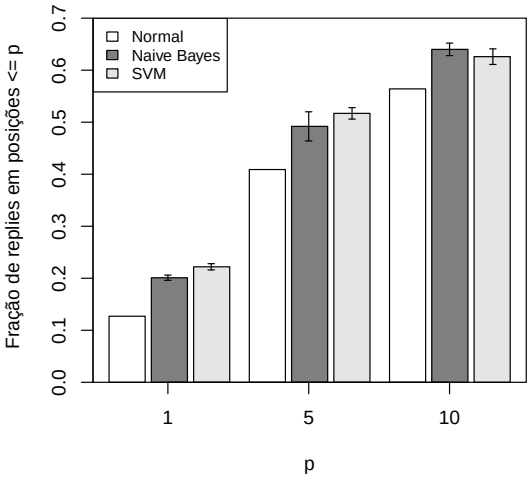
(b) Informação de Replies



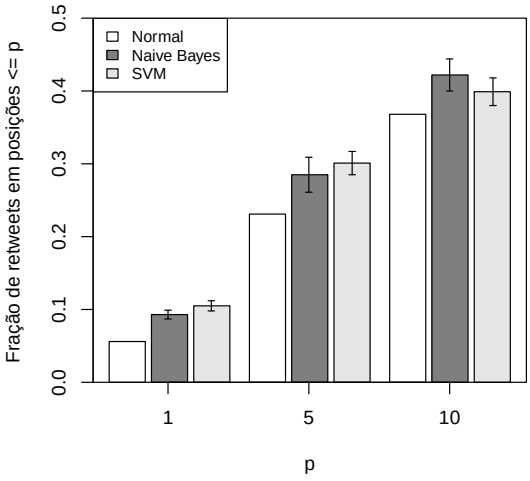
(c) Informação de Retweets



(d) Informação de Retweets



(e) Informação de Replies e Retweets



(f) Informação de Replies e Retweets

Figura 5.5: Frações de Replies e Retweets nas p primeiras posições do timeline. Barras de erro representam intervalos de confiança de 95%.

visualização, apenas os valores extremos (máximo e mínimo) são apresentados. Pode-se perceber que para ambos os algoritmos (NB e SVM) a variação (diferença entre máximo e mínimo) é pequena, indicando que nenhuma das escolha do parâmetro T_{OFF}, no intervalo citado impactou significativa e negativamente a qualidade dos resultados obtidos.

Tabela 5.1: Frações de Replies e Retweets para diferentes valores de T_{OFF}

		SVM		NB	
	p	Mínimo	Máximo	Mínimo	Máximo
Reply	1	0.21	0.22	0.18	0.20
	5	0.50	0.54	0.47	0.51
	10	0.62	0.67	0.61	0.65
Retweet	1	0.07	0.08	0.07	0.08
	5	0.23	0.28	0.23	0.27
	10	0.32	0.38	0.34	0.39

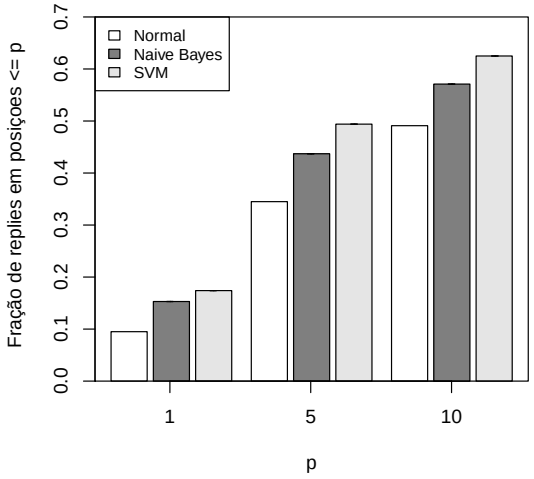
5.2.4 Usuários Ativos e Passivos

O objetivo desta seção é mostrar que a metodologia proposta é capaz de trabalhar adequadamente em conjuntos de usuários com diferentes padrões de atividades. Para este fim, foi conduzido um experimento de acordo com os moldes das duas seções anteriores, mas considerando dois conjuntos de U: o primeiro é constituído por usuários ativos, aqueles que passam mais tempo no estado ON, e o segundo, por usuários que passam menos tempo neste estado. Para dividir os usuários nestes dois conjuntos foi definida a variável $R_{ON}(u)$ como sendo a fração de tempo que o usuário u fica no estado ON no intervalo de tempo compreendido entre seu primeiro e último tweet da coleção de dados. Após isso, computou-se $R_{ON}(u)$ para todo $u \in U$ e considerou-se os 20% dos usuários com maiores valores de $R_{ON}(u)$ como sendo os usuários do conjunto ativo e os demais, como sendo os do conjunto passivo.

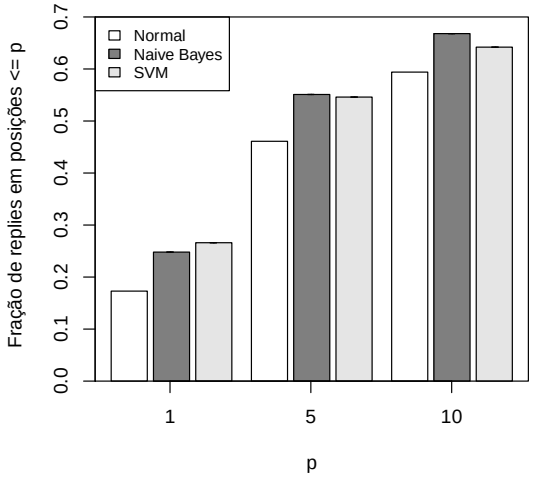
Foram extraídas duas amostras aleatórias com 2000 usuários de cada conjunto, as quais foram utilizadas no processo de simulação descrito no Algoritmo 2. É importante observar que em ambos os casos os algoritmos foram treinados com informações de replies e retweets obtidas de uma das sub amostras utilizadas na Seção 5.2.2 (tal amostra foi escolhida aleatoriamente).

A Figura 5.6 mostra os resultados deste experimento. O primeiro ponto interessante é que os usuários do conjunto passivo tendem a interagir mais com tweets próximo do topo do timeline do que os do conjunto ativo. Possivelmente, a razão deste fato é que usuários ativos passam mais tempo no estado ON e, desta forma, passam

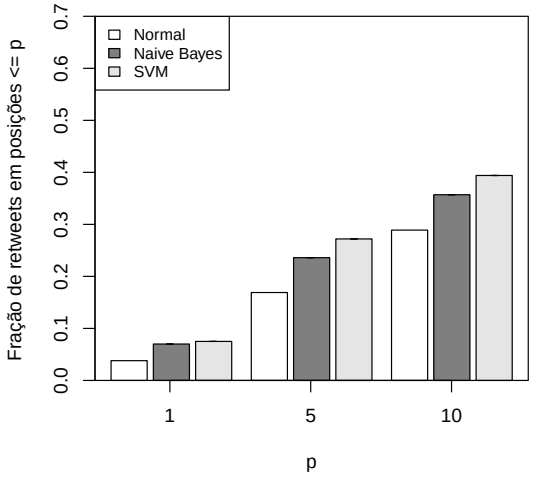
mais tempo interagindo com seus timelines e consequentemente com tweets longe do topo. Ao passo que usuários do conjunto passivo tendem a ficar menos tempo ON e portanto tendem a ter contato apenas com os tweets mais novos.



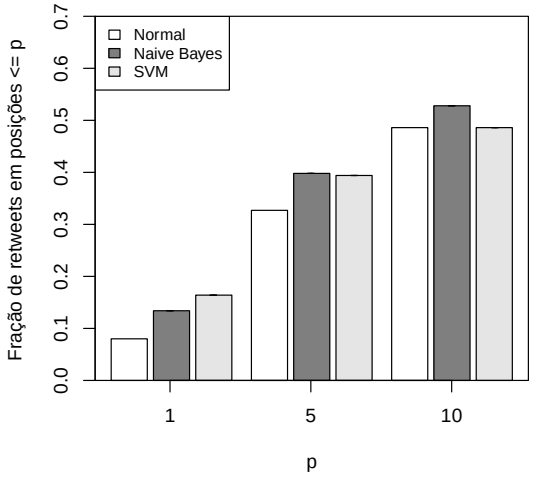
(a) Replies usuários ativos



(b) Replies usuários passivos



(c) Retweets usuários ativos



(d) Retweets usuários passivos

Figura 5.6: Comparação dos algoritmos de reorganização para dois conjuntos de atividade.

O segundo ponto é que melhorias foram obtidas para ambos os conjuntos, especialmente para usuários ativos, onde NB e SVM propiciaram bons resultados. Para o conjunto de passivos, os algoritmos também trabalharam de forma adequada, exceto para a fração de retweets quando foi utilizado o classificador SVM.

5.3 Discussões

Este capítulo apresentou uma metodologia para criar um timeline alternativo para o Twitter. A ideia consiste basicamente em utilizar algoritmos de classificação de dados para apresentar no topo do timeline os tweets com mais chances de serem respondidos ou compartilhados. Através de um estudo de simulação mostrou-se que os algoritmos propostos são eficazes e robustos em várias situações. Além disso, para cada tweet foram utilizados apenas 3 atributos os quais são simples de serem calculados, viabilizando assim a implementação para dispositivos móveis (devido a restrições de memória, bateria e processamento).

Capítulo 6

Conclusões e Trabalhos Futuros

Nesta dissertação abordou-se o problema de entender como os usuários do Twitter interagem com as mensagens em seus timelines e os ouros usuários da rede. Através de um estudo de caracterização extensivo mostrou-se a importância do problema e um conjunto de características importantes para lidar com ele. Foi mostrado que em geral os usuários preferem interagir com os tweets mais novos, com usuários que já interagiram previamente e com usuários que tenham uma baixa taxa de envio de tweets, ou seja, aqueles que não enchem seus timelines. Além disso, observou-se que algumas características textuais dos tweets, tais como, o seu número de caracteres e a presença de *mentions*, *hashtags* e URLs também afetam padrões de interações, mas de forma diferente quando são considerados replies e retweets.

Essas descobertas motivaram o projeto de um algoritmo para mudar a forma com que tweets são apresentados nos timelines dos usuários. A metodologia proposta é baseada em dois algoritmos de aprendizado de máquina, os quais mostraram melhorias significativas na reordenação do timeline através de um estudo de simulação. Os melhores resultados foram obtidos ao utilizar informações de replies e retweets para treinar os classificadores, onde foram obtidas boas taxas de mensagens respondidas e compartilhadas nas p primeiras posições dos timelines modificados. Além disso, treinando o algoritmo com informações de uma amostra aleatória obtida de todos os usuários, foi mostrado que a metodologia foi capaz de dar bons resultados para usuários em conjuntos com padrões de interações distintos, mostrando a robustez dos métodos.

Esses resultados são uma importante contribuição, uma vez que dão origem a uma outra opção de interface para usuários do Twitter, a qual pode ser especialmente interessante para os que fazem uso de dispositivos portáteis com telas de tamanho reduzido. É importante salientar que no processo de reorganização foram utilizados apenas três atributos simples e fáceis de calcular. Este fato faz esta abordagem ainda mais in-

teressante para dispositivos portáteis, dado suas restrições de memória, processamento e energia.

Este trabalho abre vertentes para alguns estudos futuros, entre os quais podem ser citados:

- ***Estudo com usuários***: os estudos de simulação propiciaram evidências de que os algoritmos de reorganização de timeline funcionam adequadamente. No entanto, um estudo com usuários através da construção de um protótipo, permitiria uma análise mais profunda dos algoritmos e também do impacto que estes teriam sobre o sistema como um todo;

- ***Investigação de mais atributos*** até o momento foram utilizados apenas três atributos para os algoritmos de classificação. Uma extensão imediata é investigar e utilizar outros para aumentar a fração de tweets respondidos e compartilhados nas posições inicias do timeline. Por exemplo, explorar o local de origem dos usuários e a importância da localização no processo de interação.

- ***Técnicas mais avançadas de aprendizado de máquina***: aprimorar o arcabouço de técnicas de aprendizado de máquina utilizado. Apesar dos resultados obtidos terem apresentado melhorias significativas, as técnicas podem ser melhoradas. Por exemplo, a versão do SVM utilizada representa um classificador binário, a qual claramente não é a mais adequada para o problema abordado. Investigar técnicas de *learning to rank* e combinar estas técnicas também são opções a serem consideradas. Outra vertente neste contexto é uma análise mais aprofundada dos procedimentos usados para amostragem na etapa de treinamento dos classificadores;

- ***Novas métricas de avaliação de ranking***: estudar e aprimorar a análise dos algoritmos utilizados fazendo uso de outras métricas de *ranking* além da precisão.

Referências Bibliográficas

Ahn, Y.-Y.; Han, S.; Kwak, H.; Moon, S. & Jeong, H. (2007). Analysis of topological characteristics of huge online social networking services. Em *World Wide Web Conference (WWW)*, pp. 835--844.

Aljohani, N. R.; Alahmari, S. A. & Aseere, A. M. (2011). An organized collaborative work using twitter in flood disaster. Em *ACM WebSci'11*, pp. 1--2. WebSci Conference 2011.

André, P.; Bernstein, M. & Luther, K. (2012). Who gives a tweet? Evaluating Microblog Content Value. Em *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pp. 471--474, New York, NY, USA. ACM.

Arlitt, M. (2000). Characterizing web user sessions. *SIGMETRICS Performance Evaluation Review*, 28(2):50--63.

Backstrom, L.; Boldi, P.; Rosa, M.; Ugander, J. & Vigna, S. (2011). Four degrees of separation. *CoRR*, abs/1111.4570.

Baeza-Yates, R. & Ribeiro-Neto, B. (2011). *Modern Information Retrieval*. Addison-Wesley, segunda edição.

Bakshy, E.; Hofman, J. M.; Mason, W. A. & Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. Em *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pp. 65--74, New York, NY, USA. ACM.

Bárabási, A.-L. (2005). The origin of bursts and heavy tails in humans dynamics. *Nature*, 435:207.

Benevenuto, F. (2010). *An Empirical Analysis of Interactions in Online Social Networks*. Tese de doutorado, Universidade Federal de Minas Gerais.

Benevenuto, F.; Magno, G.; Rodrigues, T. & Almeida, V. (2010). Detecting spammers on twitter. Em *Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*.

Bernstein, M. S.; Suh, B.; Hong, L.; Chen, J.; Kairam, S. & Chi, E. H. (2010). Eddi: interactive topic-based browsing of social status streams. Em *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pp. 303--312, New York, NY, USA. ACM.

Blei, D. M.; Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107--117.

Brlanchard, P. & Hongler, M.-O. (2007). Modeling human activity in the spirit of barabasi's queueing systems. *Physical Review E*, 75(2 Pt 2):026102.

Brzozowski, M. J. & Romero, D. M. (2011). Who Should I Follow? Recommending People in Directed Social Networks. Em *AAAI Int'l Conference on Weblogs and Social Media (ICWSM)*.

Cha, M.; Haddadi, H.; Benevenuto, F. & Gummadi, K. P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. Em *AAAI Int'l Conference on Weblogs and Social Media (ICWSM)*.

Chang, C.-C. & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27.

Cochran, W. G. (1977). *Sampling Techniques, 3rd Edition*. John Wiley.

Comarela, G.; Crovella, M.; Almeida, V. & Benevenuto, F. (2012). Understanding Factors that Affect Response Rates in Twitter. Em *Proceedings of ACM Hypertext*, Milwaukee, WI.

Cormen, T. H.; Leiserson, C. E.; Rivest, R. L. & Stein, C. (2009). *Introduction to Algorithms*. The MIT Press, 3rd edição.

Counts, S. & Fisher, K. (2011). Taking It All In? Attention in Microblog Consumption. Em *AAAI Int'l Conference on Weblogs and Social Media (ICWSM)*.

Crane, R.; Schweitzer, F. & Sornette, D. (2010). Power law signature of media exposure in human response waiting time distributions. *Physical Review E*, 81(5):056101+.

Crovella, M. E. & Bestavros, A. (1997). Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Trans. Netw.*, 5(6):835--846.

Das Sarma, A.; Das Sarma, A.; Gollapudi, S. & Panigrahy, R. (2010). Ranking mechanisms in twitter-like forums. Em *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pp. 21--30, New York, NY, USA. ACM.

Dunbar, R. (1993). Coevolution of neocortex size, group size and language in humans. *Behavioral and Brain Sciences*, 16(4):681–735.

Ebel, H.; Mielsch, L. I. & Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Physical Review E*, 66:035103+.

Erramilli, V.; Yang, X. & Rodriguez, P. (2011). Explore what-if scenarios with SONG: Social Network Write Generator. *CoRR*, abs/1102.0699.

Galuba, W.; Aberer, K.; Chakraborty, D.; Despotovic, Z. & Kellerer, W. (2010). Outtweeting the Twitterers - Predicting Information Cascades in Microblogs. Em *3rd Workshop on Online Social Networks (WOSN'10)*.

Gomide, J.; Veloso, A.; Jr., W. M.; Almeida, V.; Benevenuto, F.; Ferraz, F. & Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of twitter. Em *ACM SIGWEB Web Science Conference (WebSci)*.

Gutenberg, G. & Richter, R. F. (1944). Frequency of earthquakes in California. *Bulletin of the Seismological Society of America*, 34:185–188.

Hastie, T.; Tibshirani, R. & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer-Verlag.

Hong, L.; Dan, O. & Davison, B. D. (2011). Predicting popular messages in twitter. Em *WWW (Companion Volume)*, pp. 57–58.

Hopcroft, J.; Lou, T. & Tang, J. (2011). Who will follow you back? reciprocal relationship prediction. Em *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pp. 1137--1146, New York, NY, USA. ACM.

Hsu, C.-W. & Lin, C.-J. (1999). A simple decomposition method for support vector machines. *IEEE Transactions on Neural Networks*, 12:291--314.

Huberman, B. A. & Adamic, L. A. (1999). The nature of markets in the world wide web. Computing in Economics and Finance 1999 521, Society for Computational Economics.

Huberman, B. A.; Romero, D. M. & Wu, F. (2008). Social networks that matter: Twitter under the microscope. *ArXiv e-prints*.

Ienco, D.; Bonchi, F. & Castillo, C. (2010). The meme ranking problem: Maximizing microblogging virality. Em *ICDM Workshops*, pp. 328–335.

Jain, R. K. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley, 1 edição.

Java, A.; Song, X.; Finin, T. & Tseng, B. (2007). Why We Twitter: Understanding Microblogging Usage and Communities. *Procedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007*.

Krishnamurthy, B.; Gill, P. & Arlitt, M. (2008). A few Chirps about Twitter. Em *Proceedings of the first Workshop on Online Social Networks*, WOSN '08, pp. 19--24, New York, NY, USA. ACM.

Kwak, H.; Lee, C.; Park, H. & Moon, S. (2010). What is Twitter, a Social Network or a News Media? Em *Int'l World Wide Web Conference (WWW)*, pp. 591--600.

Lu, E. T. & Hamilton, R. J. (1991). Avalanches of the distribution of solar flares. *Astrophysical Journal*, 380:89–92.

Malmgren, R. D.; Stouffer, D. B.; Motter, A. E. & Amaral, L. A. N. (2008). A poissonian explanation for heavy tails in e-mail communication. *PNAS*.

Menascé, D.; Almeida, V.; Fonseca, R. & Mendes, M. (1999). A methodology for workload characterization of e-commerce sites. Em *ACM conference on Electronic commerce (EC)*.

Milgram, S. (1967). The small world problem. *Psychology Today*, 2:60--67.

Mislove, A.; Marcon, M.; Gummadi, K.; Druschel, P. & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. Em *ACM SIGCOMM Conference on Internet Measurement (IMC)*, pp. 29--42.

Neukum, G. & Ivanov, B. A. (1994). Crater size distributions and impact probabilities on earth from lunar, terrestrial-planet, and asteroid cratering data. *Hazards due to comets and asteroids, Space Science Series*, pp. 359–416.

Oke, A. & Bunt, R. (2002). Hierarchical workload characterization for a busy web server. Em *Int'l Conference on Computer Performance Evaluation, Modelling Techniques and Tools (TOOLS)*.

Oliveira, J. G. & Vazquez, A. (2009). Impact of interactions on human dynamics. *Physica A*, (388):187 – 192.

Radicchi, F. (2009). Human activity in the web. *Phys. Rev. E*, 80(2):026118.

Roberts, D. C. & Turcotte, D. L. (1998). Fractality and self-organized criticality of wars. *FractalsComplex Geometry Patterns and Scaling in Nature and Society*, 6(4):351--357.

Rodrigues, T.; Benevenuto, F.; Cha, M.; Gummadi, K. P. & Almeida, V. (2011). On Word-of-Mouth Based Discovery of the Web. Em *ACM SIGCOMM Internet Measurement Conference (IMC)*.

Romero, D. M. & Kleinberg, J. M. (2010). The Directed Closure Process in Hybrid Social-Information Networks, with an Analysis of Link Formation on Twitter. Em *AAAI Int'l Conference on Weblogs and Social Media (ICWSM)*.

Romero, D. M.; Meeder, B. & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. Em *Proceedings of the 20th international conference on World wide web*, WWW '11, pp. 695--704, New York, NY, USA. ACM.

Saez-Trumper, D.; Comarela, G.; Baeza-Yates, R.; Almeida, V. & Benevenuto, F. (2012). Finding trendsetters in information networks. Em *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2012)*, Pequin, China.

Saez-Trumper, D.; Meira, W. & Almeida, V. (2011). From total hits to unique visitors model for election's forecasting. Em *ACM WebSci'11*, pp. 1--2. WebSci Conference.

Suh, B.; Hong, L.; Pirolli, P. & Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. Em *Proceedings of the IEEE Second International Conference on Social Computing (SocialCom)*, pp. 177--184.

Vazquez, A.; Barabasi, A. L.; Dezso, Z.; Goh, K. I.; Kondor, I. & Oliveira, J. G. (2005). Modeling bursts and heavy tails in human dynamics. *Phys. Rev. E*, 73(physics/0510117):036127. 19 p.

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, 393(6684):440--442.

Weng, J.; Lim, E.-P.; Jiang, J. & He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. Em *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pp. 261--270, New York, NY, USA. ACM.

Weng, L.; Flammini, A.; Vespignani, A. & Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific Reports*, 2.

Yin, D.; Hong, L.; Xiong, X. & Davison, B. D. (2011). Link formation analysis in microblogs. Em *ACM Special Interest Group on Information Retrieval (SIGIR)*, pp. 1235–1236.

Zanette, D. H. & Manrubia, S. C. (2001). Vertical transmission of culture and the distribution of family names. *Physica A: Statistical Mechanics and its Applications*, 295(1-2):1--8.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley (Reading MA).

Apêndice A

Lei de Potência

Diz-se haver uma relação de Lei de Potência quando a probabilidade de se medir um valor de alguma quantidade (variável) é inversamente proporcional a uma potência deste valor. Formalmente, considere $f(x)d(x)$ como sendo a probabilidade de uma variável aleatória X assumir valores entre x e $x + dx$ (dx é um valor infinitesimal, conhecido como diferencial de x). Então, X segue uma Lei de Potência se

$$f(x) = Cx^{-\alpha}, \tag{A.1}$$

onde C é uma constante de integração, $x > 0$ e $\alpha > 1$.

Como consequência desta definição tem-se que

$$\ln(f(x)) = -\alpha \ln(x) + \ln(C), \tag{A.2}$$

ou seja, em escala logarítmica o gráfico de $f(x)$ é uma reta com inclinação $-\alpha$. Este fato implica que a distribuição empírica de uma amostra aleatória de X será semelhante a uma reta, fazendo com que o valor do parâmetro α, comummente referido por "expoente da distribuição de Lei de Potência", possa ser estimado através de uma regressão linear simples. Apesar de intuitiva, esta estratégia nem sempre traz resultados precisos uma vez que este tipo de gráfico, em geral, tem uma grande quantidade de ruído na cauda da distribuição (devido a presença de uma grande quantidade de valores com baixa frequência).

Quando estimativas mais precisas são requeridas é comum fazer uso de $F_C(x)$, Função de Distribuição Acumulada Complementar de X (comummente denotada por CCDF, do inglês *Complementary Cumulative Distribution Function*). Esta estratégia é adotada devido a uma particularidade de $F_C(x)$. Partindo-se de sua definição, dada

por:

$$F_C(x) = P(X > x), \tag{A.3}$$

tem-se que:

$$\begin{aligned} F_C(x) &= \int_x^\infty f(x')dx' \\ &= \frac{C}{\alpha - 1} x^{-(\alpha-1)}. \end{aligned} \tag{A.4}$$

Através desta equação pode-se perceber que $F_C(x)$ também satisfaz a definição de uma relação de Lei de Potência, mas com expoente $\alpha - 1$. Este resultado implica que o gráfico de $F_C(x)$ em escala logarítmica também tem a forma de uma reta (mas agora com inclinação $-(\alpha - 1)$). O interessante é que numa distribuição empírica o ruído que ocorre na cauda no gráfico de $f(x)$ não ocorre neste caso, fazendo com que o processo de regressão linear simples possa ser realizado com mais precisão.

É importante ressaltar que o valor esperado de X só é finito se $\alpha > 2$ e a variância apenas se $\alpha > 3$. Este fato é interessante uma vez que na natureza muitos fenômenos seguem uma Lei de Potência com expoentes variando de 1 a 3, ou seja, uma variância não finita.

É importante ressaltar que variáveis com distribuições de Leis de Potência estão presentes em várias áreas do conhecimento não diretamente relacionadas com esta dissertação. Entre os contextos mais comuns estão: intensidade de terremotos [Gutenberg & Richter, 1944], tamanho de crateras na Lua [Neukum & Ivanov, 1994], explosões solares [Lu & Hamilton, 1991], arquivos de computadores [Crovella & Bestavros, 1997], guerras [Roberts & Turcotte, 1998], frequência de uso de palavras em qualquer língua [Zipf, 1949], frequência de nomes em algumas culturas [Zanette & Manrubia, 2001] e número de acessos à páginas *Web* [Huberman & Adamic, 1999].

Apêndice B

Tamanho de amostras para o cálculo de proporções

Antes de iniciar este Apêndice é importante salientar que, por simplicidade, foi utilizada a notação de Cochran [1977] e que os símbolos aqui utilizados não devem ser confundidos com os apresentados no retante da dissertação.

Considere uma população composta de N unidades, as quais podem ser classificadas em duas classes, C ou C'. Seja p a proporção de elementos da classe C entre o total e $q = 1 - p$ a proporção da classe C'. Suponha agora que seja definido o seguinte problema: Qual o menor valor de n tal que uma amostra aleatória com n elementos permita estimar p com um erro menor que d e confiança α? Em outras palavras, deseja-se encontrar o menor n tal que

$$P(|p - \hat{p}| \leq d) = \alpha, \tag{B.1}$$

onde $\hat{p}$ é a proporção amostral de elementos que pertencem a classe C. Da mesma forma, define-se $\hat{q} = 1 - \hat{p}$.

Segundo Cochran [1977] tem-se que:

$$n = \frac{\frac{t^2 \hat{p}\hat{q}}{d^2}}{1 + \frac{1}{N}\left(\frac{t^2 \hat{p}\hat{q}}{d^2} - 1\right)}, \tag{B.2}$$

onde t é a abscissa da distribuição normal padrão que corta uma área $1 - \alpha$ das caudas da densidade. Em outras palavras, t é tal que $P(-t \leq Z \leq t) = \alpha$, onde Z é uma variável aleatória de distribuição normal com média 0 e desvio padrão 1.

No caso de N muito grande, ou quando deseja-se fazer um cálculo pessimista,

considera-se o limite da equação (B.2) quando $N \to \infty$, ou seja, tem-se:

$$n = \frac{t^2 \hat{p}\hat{q}}{d^2}. \tag{B.3}$$

Agora repare que no caso comum os valores de $\hat{p}$ e $\hat{q}$ não são conhecidos. Para contornar este problema, substitui-se o produto $\hat{p}\hat{q}$ pelo valor máximo que ele pode assumir. Uma vez que $\hat{p}$ e $\hat{q}$ são proporções, tem-se $0 \leq \hat{p} \leq 1$ e $0 \leq \hat{q} \leq 1$. Logo, como $\hat{q} = 1 - \hat{p}$, $\max\{\hat{p}\hat{q}\} = \frac{1}{4}$. Portanto, uma estimativa ainda mais pessimista para o valor de n é:

$$n = \frac{t^2}{4d^2}. \tag{B.4}$$

Instanciando essa equação para a aproximação do tamanho amostral apresentada no Capítulo 5 tem-se:

- Confiança de 95%, logo $t = 1.96$;
- Erro absoluto inferior a 0.005, logo $d = 0.005$.

Assim, $n = \frac{1.96^2}{4 \times 0.005^2} = 38,415.999$. Como o tamanho de uma amostra deve ser um valor inteiro, é assumido para n o valor de 38,416.