Construção de Amostras de Dados do Twitter

Tiago Magalhães[2] e Sérgio Nunes[1,2]

[1] INESC TEC
[2] DEI, Faculdade de Engenharia, Universidade do Porto, Portugal

{ei04087,sergio.nunes}@fe.up.pt

Resumo A crescente utilização de redes sociais, de que é exemplo o Twitter, tem vindo a atrair o interesse científico, originando diversos estudos e projetos de investigação acerca dos hábitos e características das suas comunidades. Além dos estudos sociais e estatísticos, caracterizadores dessas comunidades, o Twitter serviu igualmente de base para estabelecer paralelismos entre o que aí ocorre e o que poderá refletir do mundo real. No entanto, muitos destes estudos ignoram os possíveis enviesamentos que a obtenção destes dados poderá originar, colocando em risco a validade das conclusões. O principal objetivo deste trabalho foi investigar sobre as diferenças entre as amostras obtidas através de diversos tipos de extração de dados do Twitter. Os métodos foram implementados para obter diferentes amostras representativas de cada recolha. Optou-se por estudar quatro métodos de recolha de utilizadores diferentes: recolha por IDs de utilizador gerados aleatoriamente; recolha pelos IDs de utilizador presentes nas listas de "seguidores"; recolha pelos IDs de utilizador presentes nas listas de "seguidos"; recolha dos autores dos *tweets* presentes na *sample stream*, a amostra do fluxo de dados cedida gratuitamente pelo Twitter, que corresponde a 1% da totalidade do fluxo público de *tweets*. Para cada amostra, procedeu-se a uma caracterização estatística dos seus dados, relativos aos campos mais importantes dos seus utilizadores, assim como aos mais estudados. Através da comparação entre os dados das suas amostras, foi possível observar diferenças e semelhanças entre as mesmas. Concluíu-se que a forma como retiramos dados do Twitter influência o tipo de amostra que obtemos para posterior análise. Deste modo, qualquer trabalho que tenha como base dados obtidos da rede social Twitter deve realçar todos os aspetos relativos à recolha de amostras, assim como os possíveis enviesamentos que as mesmas possam originar.

Key words: social networks, sampling techniques, research data

1 Introdução

1.1 Objetivos

O objetivo primordial foi procurar a resposta a uma pergunta simples: recolhas de dados diferentes originam amostras diferentes? Uma amostragem forte e sólida é a base que sustenta qualquer conclusão de um estudo. O presente trabalho pretende analisar as propriedades das várias amostragens obtidas, de forma a tornar claro que tipo de amostra é construída quando extraímos dados do Twitter. A informação cedida pelo Twitter acerca dos seus dados revela-se insuficiente para determinar a qualidade dos dados obtidos pelas suas APIs. Assim, através dum maior conhecimento da estrutura das coleções de dados, será mais fácil não só compreender melhor os resultados apresentados em estudos e investigações do Twitter, como também os possíveis enviesamentos que as metodologias de recolha de dados podem originar.

2 Twitter Social Network

Desde o seu lançamento oficial em Julho de 2006 até aos dias de hoje, o Twitter tornou-se numa das mais importantes redes sociais presentes na *web* [Nat]. Este serviço permite aos seus utilizadores escreverem e partilharem mensagens de estado com um grupo de seguidores, que podem variar bastante na temática, abordando desde tópicos mais rotineiros e pessoais até notícias ou reacções a eventos de importância social e internacional. Estas actualizações de estado são apelidadas de *tweets*, mensagens com o tamanho máximo de 140 caracteres. Para receber os *tweets* de outros utilizadores na sua área pessoal, um utilizador deverá "seguir" outros, num conceito que difere das relações de amizade de outras redes sociais, já que é possível um utilizador "seguir"outro, sem que o inverso se verifique. Todos os utilizadores dispõem de uma página *web* com a sua área pessoal, onde os seus próprios *tweets* e os de todos os utilizadores contidos no agrupamento denominado "seguidos" (ou seja, utilizadores cujo o utilizador tem interesse em seguir os seus *tweets*) estão dispostos numa única lista ordenada cronologicamente.

2.1 Extrair dados do Twitter

Após o estudo de diversos artigos e estudos sobre a extração de dados no Twitter, indentificaram-se as seguintes formas de extrair dados do Twitter:

- extração de dados com base nos utilizadores [KGA08,RYSG10,KLPM10];
- extração de dados com base em pesquisas de *tweets* [GCNB+,JZSC09];
- extração de dados com base na *Public Timeline* [KGA08];
- extração de dados com base no ID [Cli09]:
- sistema automático [twia];
- serviços e coleções externas [KKNG12].

No presente estudo, apenas foram implementadas metodologias baseadas na extração de dados com base nos utilizadores, extração de dados com base na *Public Timeline* e extração de dados com base no ID. Na secção 4 serão descritos com maior detalhe a sua implementação.

2.2 Qualidade da Amostragem

Nesta secção será descrito um pouco o trabalho já efetuado sobre a amostragem no Twitter quanto à qualidade dos dados recolhidos pelos métodos do Twitter. Nos trabalhos estudados, verificaram e identificaram-se as várias metodologias adotadas para a extração de dados. No entanto, começa a emergir outro tipo de análises, no qual o presente estudo se insere, que pretende verificar se as propriedades das amostragens obtidas com o recurso às metodologias utilizadas pela comunidade científica diferem entre si, e se são uma representação do uso real do Twitter. Como tal, destacam-se dois estudos em particular: "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose" [MPLC13], e "Assessing the Bias in Communication Networks Sampled from Twitter" [GBWR$^+$12].

No primeiro trabalho, é efetuada uma comparação entre os dados recolhidos através da amostra cedida gratuitamente pelo Twitter, através da *Stream API*, e os dados obtidos através do *Firehose*, para determinar se os primeiros representam, de uma forma aceitável, o fluxo real do Twitter, que em princípio o acesso pago do *Firehose* garante aos seus programadores ou investigadores. A investigação tomou como campos de análise as *hastags* mais populares presentes nas amostras, a análise de tópico dos *tweets* recolhidos, uma análise à construção da rede feita por utilizadores, ligados através da prática de *retweets* e ainda uma análise à geografia deduzida através de *tweets*.

Para a comparação das *hashtags* mais frequentes nas amostras, foi utilizado pelos investigadores o método estatístico τ de Kendall, entre listas ordenadas de *hashtags* construídas com base nas amostras recolhidas da *Stream API* e do *Firehose*. Concluí-se que, para um maior número de ocorrências, os resultados são favoráveis à qualidade de dados apresentado pela amostra de *sample*, mas para valores baixos de ocorrências, os dados da *sample* revelaram-se discordantes da amostra do *Firehose*.

Para análise de tópicos dos *tweets*, a equipa de investigação recorreu ao modelo Latent Dirichlet Allocation (LDA). Os investigadores concluíram que, com maior cobertura dos dados, a divergência entre os tópicos de ambas as amostras diminui, constatando também o inverso.

A equipa de investigadores decidiu analisar as redes de utilizadores criadas através dos *retweets* recolhidos. Analisaram tanto as medidas ao nível do nó da rede, como as medidas quanto ao nível da rede. Concluíram que, através de apenas um dia de extração, é possível identificar, em média 50-60% dos 100 utilizadores-chave. Quanto às medidas ao nível da rede, foi revelado pelo estudo uma correlação entre os índices de centralização e a quantidade de dados provinientes da *Streaming API*.

Por fim, analisaram as propriedades dos *tweets* que apresentaram *geotags* de ambas as fontes de dados. Concluiram, comparando ambas as amostras, que a *Stream API* recolhe praticamente todos os *tweets* desta natureza, consequentemente validando este método como uma base forte para os estudos que utilizam os *tweets* com *geolocation*. Concluíram, no final do estudo, que os resultados da *Stream API* dependem fortemente da cobertura de dados e do tipo de analise que o investigador pretende efetuar.

No segundo estudo, pretende-se apurar as diferenças entre amostras obtidas com a *Stream API* e com a *Search API*, assim como as redes de comunicação construídas a partir das mesmas. Para tal, decidiram filtar os *tweets* relativos às manifestações políticas decorridas em Maio, em Espanha, mais precisamente nos dias 12 e 15. A recolha decorreu durante todo o mês. Foram construídas duas redes de comunicação diferentes: uma baseada nas menções a utilizadores, outra nos seus *retweets*.

Numa primeira abordagem, os investigadores descobriram que, embora maior parte dos dados esteja presente em ambas as amostras, existem registos da amostra obtida com a *Search API* que não foram extraídos pela *Stream API*: 2.5% dos *tweets*, 1% dos utilizadores e 1.3% das *hashtags*. Todavia, a amostra correspondente à *Stream API* é maior, e, consequentemente, contém atividade (utilizadores e *tweets*) que não se encontra presente na mais pequena, obtida com a *Search API*.

Consequentemente, os investigadores retiraram duas conclusões: primeiro, que as redes formadas pelas menções são mais enviesadas para os utilizadores centrais do que as redes formadas por *retweets*, e que este enviesamento poderá estar subestimado se a amostra maior é também ela enviesada para os utilizadores centrais, quando comparado com o fluxo de *tweets* completo do Twitter.

3 Métodos da API do Twitter

Na presente secção serão descritos os métodos das APIs do Twitter [Twib] utilizados nas metodologias utilizadas neste trabalho. O método *users/lookup* aceita

API	Método	Argumentos	Resultado	Limite (15minutos)
REST API	*users/lookup*	*ids* de utilizador	utilizadores	180
	friends/ids	*id* de utilizador, cursor	*ids* de utilizador	15
	followers/ids	*id* de utilizador, cursor	*ids* de utilizador	15
Stream API	*statuses/sample*	nenhum	*tweet* e utilizador	Não se aplica

Tabela 1. Tabela descritiva dos métodos utilizados no presente trabalho.

até 100 números de *id*, devolvendo a informação respectiva a cada um dos utilizadores. Caso todos os *ids* sejam pertencentes a contas de utilizador desconhecidas, suspensas ou eliminadas, não será devolvida qualquer informação relativa a esses *ids*, mas sim um objecto de erro. Caso contrário, será devolvida a informação relativa aos utilizadores válidos.

Os métodos *friends/ids* e *followers/ids* 1 devolvem os *ids* correspondentes aos utilizadores presentes na lista de "seguidos" e "seguidores" de um dado

utilizador. Este procedimento obriga a alguns cuidados na sua implementação, já que os argumentos dos métodos, assim como as respostas, são limitados. Ambos os métodos devolvem 5 000 números de *id*.

Os métodos pertencentes à Stream API funcionam de forma díspar dos anteriormente referidos. É criado um *end-point*, onde, iterativamente, o método vai devolvendo *tweets* em tempo real.

4 Metodologias para Extração de Dados no Twitter

Foram assim escolhidos quatro métodos de extração:

- exploração das ligações entre utilizadores ("seguidos" e "seguidores");
- criação de números aleatórios para retirar o utilizador ou tweet com o número de ID correspondente;
- extração dos dados fornecidos através da Stream API.

Serão explicadas no presente secção todas as decisões tomadas na escolha dos métodos, assim como a implementação de cada um.

4.1 Exploração das ligações entre utilizadores

Este método explora, iterativamente, as listas de "seguidores" e "seguidos" pertencentes ao utilizador. Ambas as listas serão exploradas separadamente, para analisar as características pertencentes aos utilizadores recolhidos por ambos os métodos.

A primeira decisão a ser tomada centra-se na escolha do utilizador ou conjunto base de utilizadores.

No conjunto base para o *script* de amostragem genérica farão parte utilizadores populares no contexto português. Como fatores de escolha, decidiu-se por utilizadores populares que interagem com o Twitter de forma real e direta, que escrevem em português, e que além de um número razoávelmente elevado de "seguidores", também possuam um número relevante de "seguidos". Podemos consultar esses utilizadores na Tabela 2.

username	seguidores	seguidos	tweets
corpodormente	126 657	158	3 384
pedrotochas	18 840	726	3 309
havidaemmarkl	99 867	18 697	8 608
fernandoalvim	74 993	78	3 001
davidfonseca	54 855	166	5 228
brunoaleixo	46 646	34 879	548
pauloquerido	72 500	44 320	98 721

Tabela 2. Conjunto base de utilizadores.

Optou-se também, no presente estudo, não estabelecer qualquer tipo de limite associado ao número de "seguidores" ou "seguidos" de um utilizador a recolher.

A terceira decisão centra-se na inclusão ou exclusão de utilizadores com listas elevadas de "seguidores" ou "seguidos", para a composição da amostra final. Decidiu-se pela escolha mais unânime verificada na revisão bibliográfica. Decidiu-se incluir todos os utilizadores, independentemente do tamanho das suas listas de "seguidores" e de "seguidos".

O método utilizado assenta na combinação de dois métodos da API (*friends/ids* e *followers/ids*) com um terceiro (*users/lookup*).

Para controlar o número de chamadas à API, foi utilizado o campo *followers-count* e *friendscount*, que representam o número de "seguidores" e "seguidos" respectivamente. Este assume o valor de -1 para o primeiro conjunto de ids. Pode-se observar com maior detalhe a recolha de seguidores no Algoritmo 1:

```
Data: integer userid is the numeric identifier of the user being crawled
Data: integer followerscount is the number of followers of the given user
count = followerscount/5000
if count == 0 then
    followers = getFollowers(userid,-1).ids
else
    response = getFollowers(userid,-1)
    followers = response.ids
    for i ← 0 to count do
        response = getFollowers(userid,response.next_cursor)
        followers.concat(response.ids)
    end
    return followers
end
```

Algoritmo 1: Pseudo-código do Algoritmo de extração das listas dos "seguidores" de utilizador.

O método é equivalente para obter os dados relativos aos "seguidos", substituindo o método de recolha de *ids* de "seguidores" pelo método relativo à recolha de *ids* de "seguidos"previamente citado.

Ambas as recolhas decorreram num período de 48 horas. O *script* com base na lista de "seguidores" iniciou a sua execução no dia 26 de Maio de 2013 às 23 horas e 0 minutos. A recolha com base na lista de "seguidos" começou a sua extração no dia 23 às 22 horas e 0 minutos.

4.2 Números de ID aleatórios

Dada a sua simplicidade conceptual, o presente método não apresenta uma grande diversidade de escolhas na sua execução. A primeira fase passa por definir o limite numérico que um ID pode assumir. Tal como DeWitt Clinton procedeu no seu artigo [Cli09], uma maneira simples de determinar o número máximo que um ID pode assumir, passa por registar uma nova conta de Twitter, anotando o ID que a aplicação atribui. Procedeu-se ao registo de uma nova conta de Twitter, para teste. Esta foi registada a 9 de Maio de 2013, às 19 horas, 31 minutos e 11 segundos, obtendo o número de ID 1416190129. De seguida, pretendeu-se recolher a informação, se disponível, correspondente aos números de *id*. Recorremos ao método *users/lookup*.

Por fim, foi necessário guardar os dados relativos aos utilizadores válidos devolvidos pelo pedido ao servidor. O *script* deu ínicio à sua execução a 9 de Maio de 2013 às 20 horas e 0 minutos, e manteve-se activo durante um período de 2 dias.

No Algoritmo 2 podemos observar com mais detalhe como funciona o método principal do *script* de recolha por ID.

```
Data: Date Time.now returns the present time
Data: Date @timelimit represents the time limit pre-defined.
while Time.now < @timelimit do
    a = lookRandomUsers
    Database.insertUsers(a,:utilizadorporid)
end
```

Algoritmo 2: Pseudo-código do Algoritmo de recolha de utilizadores pelo campo de ID implementado no presente estudo.

4.3 Dados provenientes da Stream

Para obter a amostra geral de *tweets* fornecidos pela API, é necessário utilizar o método da API *statuses/sample*.

Para definir o *end-point*, utilizamos os recursos da biblioteca TweetStream [Int], desenvolvida pela Intridea, Inc. Este permite definir o conjunto de operação a efetuar por cada iteração do ciclo de recolha. Como pretendemos utilizá-los para análise posterior, necessitamos de guardar persistentemente os dados. Essa operação é precisamente feita a cada iteração do ciclo, inserindo cada utilizador na BD, através do método *insertUserInTable*, que recebe como argumentos o objeto correspondente ao utilizador e o nome da tabela onde os dados deverão ser inseridos.

A recolha de dados iniciou-se dia 23, às 23 horas e 12 minutos, e teve a duração de 24 horas. No esquema abaixo poderemos analisar com maior detalhe o Algoritmo 3, construído para a extração desta natureza.

```
Data: string table is the name of the table destined to record the stream
      data
TweetStream::Client.new.sample do —status—
    Database.insertUserInTable(status.user,table)
```

Algoritmo 3: Pseudo-código do Algoritmo para extracção através da Stream API.

5 Descrição das Amostras

Iremos proceder à descrição destas amostras, e para cada campo do registo, será efetuada a análise comparativa dos dados. As amostras obtidas através da exploração da lista de "seguidos" e "seguidores" serão denominadas LSeguidos e LSeguidores respetivamente, a de geração de IDs aleatórios será denominada ID e a obtida através do *Sample Stream* será designada de TSample.

Para efetuar as comparações estatísticas com base nas amostras de ID, Lista Seguidos e Lista Seguidores, decidiu-se criar uma segunda amostra normalizada, filtrando utilizadores com base na atividade quanto à sua atividade na publicação de *tweets*. Optou-se por cortar todos os utilizadores que publicaram o seu último *tweet* durante todo o mês anterior ao último dia da recolha de dados. Este corte

tem como objetivo reduzir as amostras a utilizadores que, tal como os recolhidos através dos restantes métodos, possuam um grau de atividade no Twitter recente, e, assim, excluir todas as contas que possam estar abandonadas ou de uso pouco recorrente.

Na Tabela 3 é possível consultar os tamanhos de cada amostra obtida através dos diferentes métodos implementados, assim como a percentagem das amostras normalizadas face às suas amostras originais.

Amostra	Tamanho	%
ID	1 393 332	-
IDN	164 320	14,87%
LSeguidos	647 322	-
LSeguidosN	489 510	75,62%
LSeguidores	820 863	-
LSeguidoresN	372 657	45,40%
TSample	2 474 631	-

Tabela 3. Tamanho das amostras.

6 Análise de Dados

Nas secções seguintes serão analisadas as amostras quanto aos diferentes campos caracterizadores do seu perfil e da sua atividade no *Twitter*.

6.1 Tweets

Com base na figura gráfico 1, é possível verificar que as amostras apresentam características diferentes entre si. Os valores registados pela amostra ID quanto à centralidade dos dados são muito baixos, verificando uma melhoria significativa na amostra resultante do corte. Esta amostra apresenta-se com características idênticas à de Seguidores, apesar de registar valores diferentes quanto à sua variância. As amostras SeguidoresN e Seguidos também apresentam valores próximos quanto à centralidade dos valores, apesar de registarem variâncias mais díspares. No entanto, em todas as subamostras construídas com base no corte por atividade, verifica-se um aumento generalizado, tanto nos quartis e na média, assim como na variância. Mais uma vez, a diferença dos valores que se verificam entre a amostra de LSeguidores e a sua subamostra LSeguidoresN sugere a presença de mais contas inativas ou abandonadas na amostra LSeguidores face à amostra LSeguidos.

6.2 Seguidores

Neste ponto irá ser estudado o campo relativo aos "seguidores" dos utilizadores recolhidos através dos diferentes métodos.

Trata-se de um campo onde os valores têm diferenças muito significativas entre as amostras. Como observado na análise ao campo de *tweets*, efetuado no ponto anterior, a amostra de ID apresenta valores muito baixos, dificultando a sua análise objetiva. Também subamostra IDN, apesar de um aumento significativo, apresenta valores quanto à sua centralidade e variância mais baixos que as restantes. No entanto, é curioso notar o 1º quartil da amostra ID e LSeguidores praticamente coicide (7 e 8, respetivamente), e as medianas apenas distam 47, o 3º quartil de ambas as amostras apresentam uma diferença de 568. A amostra

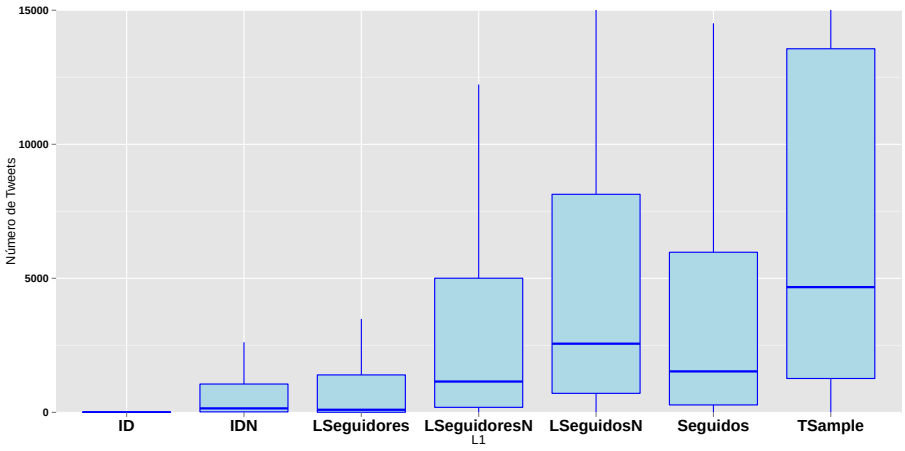

Figura 1. Diagrama de caixa e fio relativa aos valores do campo de total dos tweets emitidos por utilizador.

TSample também apresenta valores de centralidade bastante baixos, contrariando a tendência para valores mais elevados que as amostras de utilizadores com maior atividade revelam, com medianas e valores de 3º quartil na ordem dos milhares. Contudo, tanto a amostra de LSeguidos como a subamostra LSeguidosN destacam-se neste aspeto, apresentando valores bem mais elevados que as restantes quanto aos indicadores de centralidade. Mais uma vez apresentam valores superiores que a amostra LSeguidores. Também na dispersão apresentam os maiores valores.

Na Figura 2 é possível visualisar estes valores.

6.3 Seguidos

Na Figura 3 é possível visualizar estes valores representados através de um diagrama de caixa e fio. Facilmente se conclui que a amostra de ID e IDn contém valores baixos quanto à sua centralidade, fixando o seu 3º quartil ainda antes dos 250. Isto implica que 3 quartos dos utilizadores presentes nestas amostras possuem menos de 250 "seguidos". No entanto, as amostras LSeguidosN e LSeguidoresN apresentam-se bastante semelhantes quanto a este campo. Regista-se uma diferença mais acentuada dos valores quanto à sua centralidade e dispersão entre as amostras de LSeguidores e LSeguidoresN do que as registadas entre as amostras de LSeguidos e LSeguidosN. Neste campo os valores registados nas amostras de LSeguidores e LSeguidos não são tão diferentes entre si do que os registados quanto ao número total de "seguidores".

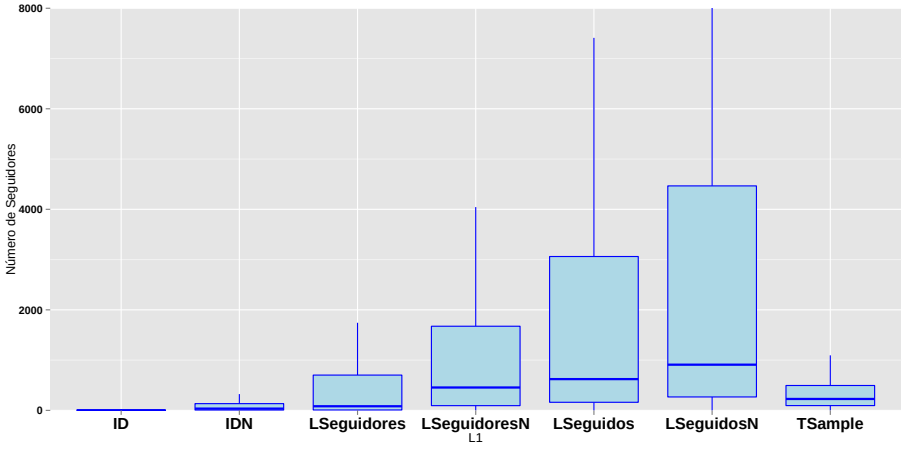

Figura 2. Diagrama de caixa e fio relativa aos valores do campo de total dos "seguidores" dos utilizadores.

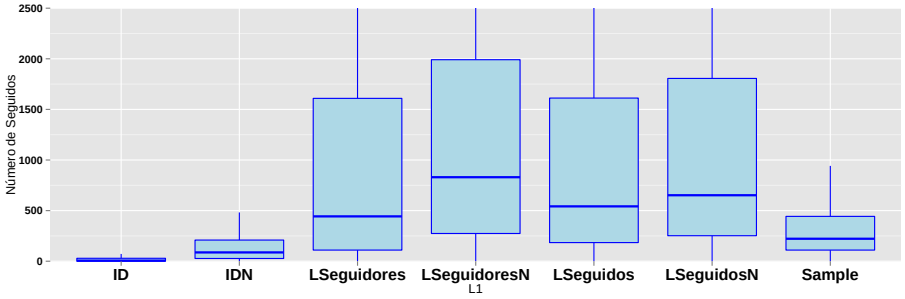

Figura 3. Diagrama de caixa e fio relativa aos valores do campo de total dos "seguidos" dos utilizadores.

6.4 Intervalos de Confiança

Nesta secção iremos calcular os intervalos de confiança das médias relativas aos campos de número total de *tweets*, total de "seguidores" e total de "seguidos", para cada amostra representativa do tipo de recolha implementado. Como grau de confiança, irá ser utilizado o valor de 95%.

Para que o cálculo dos intervalos de confiança seja robusto, é necessário que se verifiquem as seguintes condições:

- número de observações independentes;
- número de observações suficientemente grande para que a média da amostra tenda a uma distribuição Normal;

Ambas as propriedades se verificam, já que as observações em cada amostra são independentes e com ordens de grandeza na ordem dos milhares (LSeguidores

e LSeguidos) e milhões (ID e TSample). Como é possível verificar na Tabela 4, os intervalos de confiança para cada amostra apresentam valores muito diferentes para cada campo, entre as várias amostras. Estes intervalos nunca se intersectam, sendo possível concluir a sua diferença com um grau de confiança de 95%.

Amostra	IC 95%		
	Número de *tweets*	Número de "seguidores"	Número de "seguidos"
ID	616,2238 – 635,3471	97,46596 – 118,54359	95,67741 – 99,01584
LSeguidores	3 894,473 – 3 967,153	3 077,007 – 3 318,126	2722,692 – 2785,757
LSeguidos	6 723,502 – 6 810,76	17 837,54 – 18 945,96	3124,885 – 3198,022
TSample	11 662,53 – 11 742,94	1 107,292 – 1 164,512	522,2888 – 530,2142

Tabela 4. Valores dos intervalos de confiança calculados relativos a diferentes campos de cada amostra.

7 Conclusões e Trabalho Futuro

Este trabalho teve como premissa inicial responder à pergunta: técnicas de recolha de dados diferentes originam amostras diferentes? Nesta secção serão descritas as diferenças mais relevantes identificadas nas amostras, assim como as semelhanças, que confirmam algumas propriedades gerais sobre a rede social.

Através do cálculo dos intervalos de confiança referentes às médias do total de *tweets*, "seguidores" e "seguidos", é possível concluir que todos os métodos determinaram intervalos para as médias reais da população muito díspares. Os intervalos são tão díspares que nunca se intersetam. Ou seja, com um grau de confiança de 95%, nunca iremos obter amostras cuja as médias sejam identicas para os três campos analisados. Este dado evidência a grande diferença que estas amostras revelam quanto a estes campos.

Todas as amostras presentes no trabalho revelaram uma tendência para a existência de um maior número de utilizadores com mais "seguidos" do que "seguidores". Tal vem influênciar a exploração por essas listas de utilizadores, já que com menos utilizadores presentes na lista de "seguidores", são devolvidos menos números de *ids* por chamada.

Os métodos baseados na exploração das ligações de amizade dos utilizadores revelaram diferenças entre si. A Amostra de LSeguidores apresentou valores substancialmente mais altos nos campos: total de *tweets* e total de "seguidores". A combinação destes fatores sugere a presença de utilizadores com maior atividade no Twitter na amostra LSeguidos, assim como utilizadores cujos *tweets* despertam maior interesse noutros utilizadores. Uma explicação possível centra-se na maior probabilidade de existirem mais contas abandonadas ou falsas presentes nas listas de "seguidores", do que nas listas de "seguidos". Estas contas irão apresentar números reduzidos (ou nulos) de "seguidores".

A amostra através da geração de IDs aleatórios de utilizador apresenta propriedades que mais nenhum outro método de recolha estudado apresenta, sob as premissas e condicionantes definidas na secção. No entanto, apenas 14,87% da amostra corresponde a utilizadores com um grau de atividade quanto à publicação de *tweets* aceitável. A sua simplicidade de implementação e a forma

de chegar a utilizadores com poucas conexões faz com que seja uma abordagem interessante e diferente, raramente adotada pela comunidade científica.

Como trabalho futuro, não só amostras compostas por utilizadores, mas também as amostras de *tweets* podem ser alvo de estudo quanto a possíveis diferenças que possam apresentar, derivadas de diferentes metodologias e recolhas. Quanto à diferença na qualidade de amostras de utilizadores, seria do interesse científico implementar as mesmas recolhas exploratívas sob pré-condições diferentes.

Referências

[Cli09] DeWitt Clinton. Sampling Twitter. Janeiro 2009.

[GBWR+12] S. González-Bailón, N. Wang, A. Rivero, J. Borge-Holthoefer e Y. Moreno. Assessing the bias in communication networks sampled from twitter. *Available at SSRN 2185134*, 2012.

[GCNB+] Christophe Giraud-Carrier, E Shannon Neeley, Michael Dean Barnes, Kyle Prier, Joshua Heber West, Parley Cougar Hall e Carl Lee Hanson. Temporal variability of problem drinking on twitter. *Journal of Biophysical Chemistry*, 2.

[Int] Inc. Intridea. intridea / tweetstream. `https://github.com/intridea/tweetstream`. Acedido em: 27/01/2013.

[JZSC09] Bernard J Jansen, Mimi Zhang, Kate Sobel e Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.

[KGA08] Balachander Krishnamurthy, Phillipa Gill e Martin Arlitt. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks*, WOSP '08, pages 19–24, New York, NY, USA, 2008. ACM.

[KKNG12] J. Kulshrestha, F. Kooti, A. Nikravesh e K.P. Gummadi. Geographic dissection of the twitter network. In *In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

[KLPM10] H. Kwak, C. Lee, H. Park e S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[MPLC13] Fred Morstatter, Jürgen Pfeffer, Huan Liu e Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *The 7th International Conference on Weblogs and Social Media (ICWSM-13), Boston, MA.*, 2013.

[Nat] Daniel Nations. The top social networking sites. `http://webtrends.about.com/od/socialnetworking/a/social_network.htm`. Acedido em: 29/10/2012.

[RYSG10] Delip Rao, David Yarowsky, Abhishek Shreevats e Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010.

[twia] Twitter Portugal. `http://twitterportugal.com/`. Acedido em: 29/10/2012.

[Twib] Twitter. Twitter Dev Pages. `http://dev.twitter.com/pages/`. Acedido em: 29/10/2012.