A study on the characteristics of spreading news on Twitter
- The influence social media has on society

En undersökning av karakteristiken hos spridning av nyhetsartiklar på Twitter

Daniel Jonsson
Jesper Holmström

Supervisor: Niklas Carlsson
Examiner: Marcus Bendtsen
Upphovsrätt


Copyright

The publishers will keep this document online on the Internet – or its possible replacement – for a period of 25 years starting from the date of publication barring exceptional circumstances. The online availability of the document implies permanent permission for anyone to read, to download, or to print out single copies for his/hers own use and to use it unchanged for non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional upon the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility. According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement. For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its www home page http://www.ep.liu.se/.

© Daniel Jonsson
© Jesper Holmström
Students in the 5 year Information Technology program complete a semester-long software development project during their sixth semester (third year). The project is completed in mid-sized groups, and the students implement a mobile application intended to be used in a multi-actor setting, currently a search and rescue scenario. In parallel they study several topics relevant to the technical and ethical considerations in the project. The project culminates by demonstrating a working product and a written report documenting the results of the practical development process including requirements elicitation. During the final stage of the semester, students create small groups and specialize in one topic, resulting in a bachelor thesis. The current report represents the results obtained during this specialization work. Hence, the thesis should be viewed as part of a larger body of work required to pass the semester, including the conditions and requirements for a bachelor thesis.
Abstract

The spreading of news on social media is a complex process and sought-after skill in today’s society. People spreading political beliefs, marketing teams who want to make money and people who want to achieve fame are all trying to understand the best way to influence others. Many are trying to understand this complex process to limit the impact that the spreading of fake news and other misinformation may have on society. This research has gained a lot of attention recently, but no definite answers to several important questions have been found. Our main contribution is to create a methodology that allows us to collect more interesting longitudinal data, while at the same time reducing the number of calls to the used APIs. This is done by introducing a threshold that filters out links that are found to be uninteresting. We also introduce a random factor in order to eliminate and understand the bias introduced with this threshold. Our analysis of the longitudinal measurement show that there is no strong correlation between the number of followers a user has and the number of clicks a link posted by the user receives and that a link’s popularity typically is reduced significantly after its first few hours of existence. This illustrates the reactive and fast-paced nature of Twitter as a means to share information.
Acknowledgments

Firstly, we would like to thank our supervisor Niklas Carlsson for great guidance and tutoring. We would also like to thank Tobias Löfgren, Robin Ellgren, Gustav Aaro and Daniel Roos for peer-reviewing our paper, as well as Linnea Lundström and Sebastian Ragnarsson for brainstorming and discussions about different viewpoints throughout the study. Lastly we would like to thank Olav Nilsson, Filip Polbratt, Anna Vapen, Anton Forsberg and Karl Andersson who wrote the program that we used as a basis.
# Contents

**Abstract** .......................................................... iv

**Acknowledgments** .................................................. v

**Contents** ........................................................... vi

**List of Figures** ..................................................... vii

**List of Tables** ..................................................... viii

1 **Introduction** ..................................................... 1

1.1 Aim ......................................................................... 2

1.2 Research questions .................................................. 2

1.3 Contributions ......................................................... 2

1.4 Thesis outline ......................................................... 3

2 **Theory** ................................................................ 4

2.1 Related work ......................................................... 4

2.2 Fake news ............................................................. 5

2.3 Classification ......................................................... 6

3 **Method** ............................................................... 9

3.1 Data mining ........................................................... 9

3.2 Processing of tweet and Bitly blocks ................................ 10

3.3 Deriving of threshold and sets ................................. 11

3.4 Classification ........................................................ 14

3.5 Analysis of the collected data .................................... 14

3.6 Delimitations ......................................................... 14

4 **Results** ............................................................... 17

4.1 Filtration of the collected data ................................. 18

4.2 Fake versus real news ............................................ 19

4.3 Difference between categories and sets ..................... 21

4.4 Impact from a user’s of number of followers .............. 27

5 **Discussion** .......................................................... 29

5.1 Results ............................................................... 29

5.2 Method ............................................................... 30

5.3 The work in a wider context .................................... 31

6 **Conclusion** .......................................................... 32

6.1 Future work ......................................................... 32

**Bibliography** .......................................................... 34
List of Figures

3.1 Data mining flow chart . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
3.2 Block schematic over the collection of tweets and updates of Bitly links . . . . . . 10
3.3 The split of Bitly blocks at t=12 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11
3.4 Overview of calls to the Bitly API . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12
3.5 Venn-diagram of the link categories and their update interval . . . . . . . . . . . . 13
4.1 CCDF for different filtrations on the collected data at t=120 . . . . . . . . . . . . 18
4.2 CCDF for difference between BBC and Breitbart at t=120 . . . . . . . . . . . . . . 19
4.3 CCDF for difference between the remaining news sites at t=120 . . . . . . . . . . . 20
4.4 CCDF for all data sets at t=120 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21
4.5 CCDF for the random set versus all sets at t=120 . . . . . . . . . . . . . . . . . . . 21
4.6 CCDF at different timestamps for top, random and all sets . . . . . . . . . . . . . 22
4.7 Clicks at t=12 versus t=120 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23
4.8 Clicks at t=24 versus t=120 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23
4.9 Clicks added between t=24 and t=120 for top, random and all sets . . . . . . . . 24
4.10 Clicks added between t=2 and t=4 for top, random and all sets . . . . . . . . . . 25
4.11 Pearson correlation r of clicks at t and t+T . . . . . . . . . . . . . . . . . . . . . . 26
4.12 CCDF for different follower intervals at t=120 . . . . . . . . . . . . . . . . . . . . . 27
4.13 Followers versus clicks at t=120 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28
5.1 The top set with threshold set at different times . . . . . . . . . . . . . . . . . . . . 30
List of Tables

4.1 Table with statistics of all collected data .................. 19
4.2 Medium retweet rate for different news sites for our top set 20
Every day millions of people sign into their social media accounts. There they read information that will influence how they think, what they do and how they will react to events in their lives. This is comparable to how trends in clothing fashion and happenings spread in the society; extremely brisk and unpredictable. Social media and its place in peoples’ everyday life is growing rapidly. During this year’s second quarter, 75% of all internet users visited social media websites, compared to 56% one year earlier [1]. Traditional ways of getting news through papers, TV and radio broadcasts are replaced by news websites and social medias like Twitter and Facebook. According to Gottfreid et al. in 2016 62% of US adults obtained their news on social media [2]. In today’s society, this makes social media the ideal platform to spread opinions and arguments to influence others. One successful example of this is the ALS Ice Bucket Challenge in 2014 which swiftly spread all over social media and drew attention to the ALS disease. With the help from both celebrities and ordinary people who got involved, the campaign raised a total of $115 million[1]. However, this trend also makes it harder to regulate the spreading of information and mis-information.

Social media plays a central role in today’s society which makes it important to understand the dynamics of it. However, predicting which particular user or URL will generate large cascades of interest and societal impact is a hard problem with relatively unreliable results [3]. Why, how, when and by whom information is spread are all questions that needs to be explored to get a better understanding of these dynamics. However, as Bakshy’s study [3] notes, this is likely to be impossible to fully achieve. It is nevertheless knowledge that is sought after by countless people, from marketing teams and celebrities to non-famous people. With social media having such a big impact on society some people will also try to find ways to exploit this for their own winning. This is often done by spreading incorrect, biased information or so-called fake news. Unfortunately, this is often done in a way that makes it hard to identify and distinguish it from legitimate information.

Due to how easily we are influenced by the (mis)information spread over these media - consciously or unconsciously - it is important that what we read and believe in actually is true. The concept of fake news has grown in popularity recently and we can see instances where the spreading of fake news has even influenced meaningful worldwide events. The most known and maybe most controversial is how social media affected the US president

1.1 Aim

Because of the rapid growth of social media in human society and because of how easily we are influenced by what is said on social media it is undoubtedly an area of expertise worth examining. In order to understand what is spread on social media and what aspects and factors that affect this we need examine and understand popularity dynamics such as click behaviour, news cycles and classification of news. Our main contributions with our thesis is a longitudinal measurement of popularity dynamics and to examine a way of creating a threshold. This threshold is a never before used methodology of finding links that are considered interesting and worth following more frequently. We also look at how an article being classified as real or fake affect the spreading of a news article.

1.2 Research questions

The following questions are meant to be researched and answered in this thesis

1. What aspects and factors affect what is spread on social media and to what extent?
2. Is it possible to observe any patterns in the spreading with social media?
3. How well can the number of clicks at different times in the collection period determine if a link is interesting or not?

1.3 Contributions

In this thesis we extended a measurement framework to provide more careful selection of which tweets to collect information about and then used the enhanced tool set to study what factors affect the spreading of news on social media, most notably on Twitter. We use the Twitter API to collect 10.8 million tweets containing a Bitly link and collected a total of 1,877,045 unique Bitly links. Of these, 1,481 links are collected that correspond to the news sites of interest. We use the Bitly API to periodically retrieve the number of clicks that each link to an article has gotten over time, for a total of 5 days. The Bitly API was chosen because it is the only service that is being used on Twitter to such a large extent where it is possible to retrieve data about the number of times a link has been clicked. A threshold is introduced to filter out the most "interesting" Bitly links and update the number of clicks for these more often. The term interesting here is defined as links with a number of clicks over the derived threshold, explained in Section 3.3. This filtering was introduced to better utilize the Bitly API which uses a rate limit that limits the amount of allowed API calls to 1000 per hour. By focusing on the most popular tweets (and a random sample set for comparison), we can collect more data and get more reliable results. We also introduce a random set of links that we follow and update in the same way. This is to eliminate the bias that is introduced by stating that the most interesting links are the ones with clicks over the threshold.

The classifier used in this thesis project is written by last years bachelor student which is a simple classifier with fairly unreliable results. The main focus of this thesis is not classifying and therefore work will not be put in to improve this classifier.

---

This study results in several key observations worth noting. Our classifier classify considerably more links as real/objective instead of fake/biased. This may depend on the simplicity of the classifier or that more sites in our set of news sites posts articles that more often are classified as objective than biased.

It is also observed that the gain in clicks for an article early in its existence mimic the relative gain later in its existence, meaning that it is possible to know if a link will get popular early on. This implies that our threshold could be set at an earlier time than $t=12$.

The results also show that the filtrated random set follows a similar pattern in click behaviour to the set with all links, which is expected since a link has the same probability of being labeled as random regardless of its number of clicks. This indicate that our results are scalable to a bigger scope in the terms of displaying the characteristics of spreading news on social media.

Focusing on the people posting the tweet, the so called tweeters, we study the longlivness of a tweet and what aspects that affects this. A name for the most influential people used through out related works is “influentials” [5] or “true influencers” [6]. We investigate how much influence these so-called influencers have on others and how this is earned and attained. This is combined with the behaviour of spreading of fake news, which in comparison to just counting the clicks of an article link gives a much better understanding on the sharing dynamics of news on social media. Another core component that is used in this study is the set of links that is randomly selected to eliminate bias.

1.4 Thesis outline

The remainder of this thesis is structured as follows. First, we provide some background to the topic and previous work made in the area we are exploring. Afterwards we thoroughly explain our method and our developed tool and how the work was carried out. Results of our work is afterwards presented and later both the results and method of use is discussed.
2 Theory

2.1 Related work

Influence of users and their following

Several studies have been done on how people interact and influence each other, and more importantly for this work; how they interact and influence each other on social media. Both by users who want to influence others, most commonly for their own gain, but also by companies for viral marketing. Cha et al. present an in-depth study of the dynamic in user influence on social media [5]. This investigation resulted in three main interesting observations:

- Being a popular tweeter does not necessarily lead to more influence in the form of retweets or clicks.
- Tweeters with a large popularity can over multiple fields hold influence.
- To influence on social media you have to get popular and to achieve this you have to make an effort.

To try interpret how and why Cha et al. came to these conclusions we need to look at the motivation. Surprisingly, there is according to Cha et al. no direct correlation between the number of followers and retweets. Instead people who interact and stay active on Twitter by retweeting, commenting and converse with other users are the ones who receives more retweets and hence influence others on social media. Having a larger following certainly helps with reaching out to a bigger audience and having the chance to influence others, but also staying active by sharing, posting and interacting is a winning strategy. By collecting the number of followers of a tweeter this is an aspect we are able to study and analyze.

Looking at the second observation stated by Cha et al. you would also assume that focusing on a single topic would imply more influence, but the most influential users often broaden their area of focus to consequently also expand their influence.

Being influential and popular does not occur by chance either, but instead by committing to it according to the study. Staying active and interactive increases the chance to grow in popularity. It is also beneficial to not go with the stream by making generic tweets but instead stating creative and insightful opinions, preferably within a single narrow topic. As stated above doing this is more important than having a large following in order to influence others,
which is another thing that our method permits to follow. It can be observed in the results that a user with less followers than another may get more clicks on their link and therefore influence to a greater extent.

As Cha et al. states, it is possible to achieve great exposure for a product or beliefs by targeting the most influential people on social media with an already dedicated audience, for a small cost. This can be seen very clearly being used by marketing teams in today’s society where product placements can be seen all over celebrities’ social media pages as an efficient marketing method.

The difference in our conducted study is that we look at links to news sites and their spreading patterns. Cha et al. only looks at a tweets activity, as we also do but to a lesser degree. The articles are also classified as real or fake news, enabling us to examine how this aspect affect the influence and spreading dynamic.

News cycles and share dynamics

In a recent study by Gabielkov et al. which has a very similar data mining method, they concluded an unbiased study by focusing on patterns and phenomenon associated with news sharing dynamics on social media. It is discovered that a link can generate considerable amounts of clicks several days after it is posted, in sharp contrast to how a tweets activity is essentially concentrated around the first hours of its existence [6]. From their data set they try to predict the future clicks on a URL at the end of the day. They discovered that following clicks on the URL instead of number of retweets is much more accurate. When using data on the number of clicks obtained after following for four hours they receive an Pearson correlation between the predicted value and the actual value of 0.87 compared to 0.65 when using retweets. This supports why we in this thesis will follow the Bitly links activity closely and not the tweets. The long-lived activity of the Bitly links is also taken into consideration when choosing how long we follow the links and update their clicks.

An et al. presents a study on how the media landscape has changed with society moving from traditional media sources to the more global and interactive social media [7]. This paradigm shift in media journalism has lead to journalists having more freedom to actively interact with their following. An et al. also suggest that the followers themselves have a much greater opportunity to boost their influence on others, or even in some cases being able to influence at all.

The difference in our study from the previously mentioned studies is that we investigate how news being classified as real or fake news have an impact on the paradigm shift in media journalism and how this affects for how long links generate clicks. Our work is also heavily focused around our implementation of the threshold which is a new unique way of deciding if a link is interesting. By doing this, in terms of the used APIs’ rate limits, more interesting data can be collected. We also conduct our longitudinal study for 7 days while Gabielkov et al. collect links for a month, however we follow and update clicks for 120 hours while Gabielkov only update for 24 hours.

2.2 Fake news

What fake news really is depends on who is using the term. Different people who use the term defines it in different ways, but what they all have in common is that the news in some way are not real or not accurate. In an article in the magazine Science, the author Lazer D. et al. define the term as: “Fake news is fabricated information that mimics news media content in form but not in organizational process or intent” [8]. They write that fake news overlaps with other types of “information disorders”, for example information that has the purpose to mislead people. In a 2017 paper from H. Allcott et al where they define fake news as "to be news articles that are intentionally and verifiable false, and could mislead readers" and their study focus mostly on articles that has political content with a connection to the American
2.3 Classification

Text classification is one of the most common tasks in machine learning, a well known technique which rose massively in popularity over the last years. There are several different methods to classify a text and also several different categories to classify a document as. All these implementations, also called classifiers, have different pros and cons in everything from predictability to execution time. Since the classifier we use is written by last years bachelor students we will only use it as a tool and will not try to improve it. The following sections is only for enlightenment and a deeper understanding of the classifier.

Bayes' theorem

Multiple of these classifiers are developed from the well known theorem in probability statistics, Bayes' theorem. Bayes' theorem describes the conditional probability $P(A|B)$ - the likelihood of event $A$ to occur given that $B$ is true - which is based on three things:

- $P(B|A)$ which describes the probability that $B$ occurs if $A$ is true
- $P(A)$ which describes the probability that $A$ occurs
• \( P(B) \) which describes the probability that \( B \) occurs

These four probabilities give us Bayes’ theorem:

\[
P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)},
\]

which is a formula considered essential in probability theory.

The Naive Bayes classifier

The simplest classifier utilizing Bayes’ theorem is the Naive Bayes classifier [12]. The classifier used in our study is also a variant of the Naive Bayes classifier. This classifier is as its name states, naive, meaning that it assumes all the features are conditionally independent in a text, the “Naive Bayes assumption” [13]. For example, if considering a vegetable that is oval, green and roughly 20 centimeters long your first guess would probably be a cucumber. Instead of taking all the features that characterize a cucumber to account to figure out the answer a Naive Bayes classifier will instead - regardless of any possible correlation between them - independently take the features into account.

Even though this assumption seldom is correct the different variations of the Naive Bayes classifiers is known for performing well most of the time. One of the most considerable reason why the Naive Bayes model is used in classifiers is its time complexity. All the model has to compute is frequency of every word in every class, which means that both training the classifier and classifying a text has optimal time complexity [14].

There are several traits or weaknesses with this model which can critically “misinterpreted” the calculations and therefore alternate the results. As stated above there are various variations of the Naive Bayes Classifier with different proposed solutions to these factors, some explained below. These solutions obviously make the classifier more complex which affects the time complexity.

Smoothing

In Equation 2.1 it is shown that the Naive Bayes classifier calculates the probability by multiplying the probabilities \( P(B|A) \) and \( P(A) \). This means that if one of these is zero the probability for the whole class can become zero even if it shouldn’t. This could be a consequence of a classifier not trained with enough data and therefore calculates the probability of a word occurring to zero. This trait, which can describe as a weakness of the model, is partly eliminated by using a smoothed estimation to calculate \( P(A|B) \) as shown below:

\[
P(A_i|B) = \frac{N_i + \alpha}{N + \alpha n'},
\]

where \( N_i \) is the number of times the word \( i \) occurs and \( N \) is total amount of word occurrences. \( \alpha \) is a so called smoothing variable which you can set to an appropriate value. If \( \alpha=0 \) there is no added smoothing and if \( \alpha=1 \) this technique is called add-one smoothing.

Stop words and low frequency words

There are a few words that frequently are used in our language and others that almost never are used. Both in terms of time complexity and getting a proper result these words can overall excessively affect the classifier. Implementations are therefore commonly introduced to limit and tokenize the data by removing low and/or high frequently used words. Words that occur often, so called stop words, is put in a stop list and not put in to the calculations when classifying. The same can be done by removing words that occur only once, so called stemming and thus reduce the size of the corpus of words.
2.3. Classification

Bag of words

Previously we have only looked at one word independently and its number of occurrences to classify a document, also called unigram. Another similar simplifying model is Bag of words, in which you put multiple words or a sentence in a “bag” to classify the text. This means that the model ignores word ordering and grammar and takes count of frequency. This is as mentioned above a simplifying representation of the data, meaning it has good time complexity but may not be the most accurate model in most cases [15]. For example, by implementing unigram or bag of words in a model “Peter likes pizza but hates potatoes” and “Peter likes potatoes but hates pizza” would amount to the same result. In this representation it is not taken into account what food Peter likes and what food he hates.

N-grams

Instead of just looking at one word and its occurrences you can instead look at multiple words and the occurrences of their combination of ordering. These models are called N-grams where the N stands for the number of words taken into account, worth noting that this N differentiates from the N mentioned in the Section [Smoothing]. This means that a bigram is a pair of two words and trigram is three and so on. Conceptually, unigram as discussed above is a N-gram model where N=1.

If we again look at the example presented in the Section [Bag of words] by instead using the bigram model we would be able to distinguish which dish Peter likes and which he hates. For example, the first sentence would be stored as tuples by the bigram model as following:

- {"Peter, likes"}
- {"likes, pizza"}
- {"pizza, but"}
- {"but, hates"}
- {"hates, potatoes"}

As you can see the combination of liking pizza and hating potatoes is saved by this model, meaning it is possible to include the correlation between the verb and noun in the classification. One drawback with N-gram models is time complexity. Instead of just looking at one word, looking at tuples or sets to calculate instead of just looking at one word leads to more calculations. Another drawback is the increasing amount of training data needed as N grows larger. For it to result in correct data when introducing tuples or sets in the classification it is also necessary for the classifier to have seen these tuples or sets earlier to be able to have a correlating probability [16]. Simply put, by using a more complex N-gram model the classifier needs to train on more data in order to be able to evaluate and classify the data for a factual result. This means that the time it takes to train the model increase as well.

Cumulative distribution function

As a statistical tool in this paper complementary cumulative distribution function CCDF will be used. The function is the complementary function to the cumulative distribution function CDF [17]. The CDF, \( F(x) \) gives the probability for \( X > x \). The CDF for a continuous random variable \( X \) is defined as:

\[
CDF(x) = F(x) = \text{Prob}[X > x] = \int_{-\infty}^{x} f(y)dy.
\]

The CCDF is defined as the complement to CDF:

\[
CCDF(x) = \text{Prob}[X < x] = 1 - CDF(x).
\]

The CCDF will be used to show the probability that a Bitly link will have more than \( x \) clicks at a certain point in time.
3 Method

Our method is an extension of a previous bachelors thesis data collection framework, *Counting Clicks on Twitter* [18]. Our contribution to the methodology is to extend and improve the framework by making it possible to gather and follow data for a longer period of time without exceeding the rate limits of used APIs. We define the concept *follow* as updating a Bitly link’s number of clicks. This is done by the use of a threshold which decides how frequently a link’s number of clicks should be updated during our following period of 5 days, 120 hours. We also introduce a set of randomly selected links that we update more frequently, to eliminate bias as explained later in Section 3.3. The following sections will introduce and explain our data mining process and calculations to achieve a longitudinal data collection.

3.1 Data mining

![Data mining flow chart](image)

**Figure 3.1**: Data mining flow chart
The first step in the data mining algorithm is as shown in Figure 3.1 collecting tweets. Tweets are collected in periods of 20 minutes by using the Twitter streaming API and saved in what is called a "tweet block". The streaming API collect tweets posted on Twitter in real-time and due to the limitations of the Twitter API, at most 1% of the tweets including both original tweets and retweets are collected. The collected tweets are only tweets containing Bitly links and the 1% limit is more than enough for our collection. During our collection period this rate limit was never exceeded meaning no tweets containing a Bitly link was missed and therefore not collected. Blocks are gathered 504 times which amounts to a total of 7 days. The tweets are stored in JSON format and all of the tweets are saved in a .txt file.

3.2 Processing of tweet and Bitly blocks

When a tweets block is completed each tweet is scanned for Bitly URLs. Bitly is a link managing platform used for link shortening and the Bitly API has a rate limit. This limit is 100 calls per-minute or 1,000 calls per-hour and resets on every hour and simultaneously accepts five concurrent connections. These limits is per IP-address preventing us from making more accounts. The Bitly API is used to gather the number of clicks gained for the followed links. Since Twitter has a maximum length of 280 characters per tweet, link shorteners are popular among tweeters and commonly used.

The next step is to expand the Bitly URLs leading to the news article. Because of the rate limits only unique Bitly links are expanded, in other words we only expand the same link once. The expanded URLs are crosschecked against a set of selected newspaper websites. This set contains news sites that are known to publish objective accurate news or news sites that are known to publish biased or fake news. The news sites in this set was identified in a Buzzfeed news analysis [19].

![Block schematic over the collection of tweets and updates of Bitly links](image)

**Figure 3.2**: Block schematic over the collection of tweets and updates of Bitly links

---

2. https://bitly.com/pages/about
3.3 Deriving of threshold and sets

For the links that pointed to the set of the selected news sites, clicks on the Bitly links were collected from the Bitly API and from the tweets data the number of followers and retweets was collected for each Bitly link. This set of data is what will be called a completed "Bitly block". These blocks will be followed for a total time of 120 hours (5 days) and the clicks will be updated every 2 and 24 hours, as shown in Figure 3.2 and 3.3. The update interval will be decided from our threshold and if the link is labeled as random. This will be explained in the next section.

3.3 Deriving of threshold and sets

To reduce the number of calls to the Bitly API for updating the number of clicks on a link a threshold is implemented at the twelfth hour, \( t=12 \), the sixth update. This time was chosen because according to our pre-study it was observed that a link’s popularity flattens around this time. This threshold checks the amount of clicks at that current time and if the click count is above the threshold the links are moved to the 2 hour update interval, otherwise to the 24 hour update interval. The 2 hour update interval will also contain a random set of Bitly links. The random set will contain Bitly links that independently of the amount of clicks at \( t=12 \) will be updated every second hour.

In order to know what percentage of Bitly links that needs to be moved to the 24 hour block at \( t=12 \) to not exceed the Bitly API rate limit it is needed to derive the maximum amount of calls during an hour. Therefore, the following variables are introduced:

- \( p \) = “percent of Bitly links moved to the 24 hour update”
- \( a \) = “average of Bitly links in a Bitly block”
- \( n \) = “number of Bitly blocks starting every hour”

These variables will be set according to a data set of Bitly links gathered for 48 hours in order to calculate \( p \) and the optimal threshold.
3.3. Deriving of threshold and sets

Figure 3.4: Overview of calls to the Bitly API

The height of Figure 3.4 represents number of Bitly links required to update at the corresponding time. From \( t=120 \) to \( t=168 \) the largest amount of Bitly links that needs to update. This is because Bitly links are being followed and updated for 120 hours, and in interval \( t=120 \) to \( t=168 \) updates of clicks are being started and stopped at a constant rate.

We chose to calculate the amount of calls during the period \( t=120 \) to \( t=121 \), marked with dashed lines in Figure 3.4. In this interval 3 types of Bitly links have to update, all marked in the Figure 3.4. Links from the interval \( t=0 \) to \( t=108 \) which have been followed for more than 12 hours and have been split into i) 2 hour or ii) 24 hour update intervals and iii) links from the interval \( t=108 \) to \( t=120 \) which have not yet split. Calls from links in i) and ii) are given by the following expression where the numbers 54 and 5 are the number of Bitly blocks for links in the 2h respectively 24h update interval that will update in \( t=120 \) to \( t=121 \):

i) \( n \cdot 54 \cdot a(1 - p) \)

ii) \( n \cdot 5 \cdot a \cdot p \)

Links in iii) update every 2 hours and the number 6 is the number of Bitly blocks that will update its links in the interval \( t=120 \) to \( t=121 \). The number of calls for iii) is given by the following expression:

iii) \( n \cdot 6 \cdot a \)

As stated above the rate limit on the Bitly API is 1,000 calls per hour. This gives us the following equation when we add i),ii) and iii) together for the total number of Bitly calls from \( t=120 \) to \( t=121 \):

\[
n \cdot 54 \cdot a(1 - p) + n \cdot 5 \cdot a \cdot p + n \cdot 6 \cdot a < 1000
\]

The previously mentioned data set contained 1149 number of Bitly links and how many clicks each link had at \( t=12 \). This gives us an average of \( a=8 \) collected links per 20-minute Bitly block.
3.3. Deriving of threshold and sets

and \( n=3 \) collected blocks per hour. This gives \( p \geq 0.37 \). This represents the lower boundary for what percentage of the links needed to be moved to the 24 hour update interval in order to not exceed the Bitly API rate limit.

Though it would be ill-advised to pick this lower bound as our percentage to move to the 24 hour interval because of the immense variation in the number of Bitly-links collected in each block. These stochastic processes typically contain bursts and follow heavy tail distributions. Therefore a safety margin is needed which will be calculated later.

### Calculating threshold with the random set

With the previously mentioned random set it is now possible to calculate the final threshold. The purpose of this random set is to analyze the bias that occurs when it is decided that links with a number of clicks over the threshold at \( t=12 \) is more interesting than links below the threshold. This means that links that are found uninteresting because they have clicks under our threshold at \( t=12 \) but later in their existence gain huge amount of popularity and clicks could be missed. To analyze the bias we need to have a set which has statistics about a links clicks at the same times as the top links. Therefore we emanated from balancing the size of the set of links with clicks above the threshold and the set of links labeled as random, allowing us to observe and analyze the bias in the results. These randomly selected links will be updated every two hours, regardless of number of clicks.

![Venn-diagram of the link categories and their update interval](image)

**Figure 3.5:** Venn-diagram of the link categories and their update interval

As shown in Figure 3.5 there are now four different categories to sort all links in: 1) top and not random, 2) top and random, 3) bottom and random and 4) bottom and not random. Where links in 1)-3) will be put in the 2 hour update interval and links in 4) in the 24 hour. Top is the Bitly links with a number of clicks above the threshold and bottom is below.

As mentioned above the top and random set should be of equal size and to calculate the proportion of links to have in the 2 hour update interval the following equations are used:

\[
P(\text{top}) = P(\text{random})
\]

\[
1 - p = P(\text{top}) + (1 - P(\text{top})) \cdot P(\text{random})
\]

\[
1 - p = P(\text{top}) + (1 - P(\text{top})) \cdot P(\text{top})
\]
Where \( P(\text{top}) \) is the probability that a link is labeled as \textit{top} and \( P(\text{random}) \) is probability that a link is labeled as \textit{random} and as previously \( p \) is the proportion of links to be moved over to the 24 hour interval.

By choosing \( P(\text{random}) = P(\text{top}) = 0.25 \) even sets are obtained and the splitting of links between the categories of links become clear. This gives \( p = 0.5625 \), which gives a proportion that is greater that our lower boundary which greatly decreases the chance of exceeding the Bitly API rate limit. The number is not much larger than the lower boundary which means that not an excessive amount of links are moved and therefore more data is attained.

Because 25% of the links with clicks under the threshold now are defined as \textit{random} and put in the 2h update interval it is needed to inflate the threshold to compensate. To balance the size of \textit{top} and \textit{random} it is derived that the threshold should filter out 25% of all links to the \textit{top} category. This means that our threshold should according to our previously mentioned collected data set be set near 435, so the final threshold is set to 450. This means that links that aren’t labeled as \textit{random} and have less clicks than the threshold should be moved to the 24 hour update interval.

This reduces the total number of calls to the Bitly API per link from 61 calls for a link in the 2 hour update interval to 12 calls for a link in the 24 hour update interval. With the aim to move 56.25% of the links by implementing the threshold it will reduce the number of calls by 45%.

### 3.4 Classification

When a Bitly block has reached the final update at \( t=120 \) the news article is classified as real or fake news. This is done by a machine learned Naive Bayes classifier. The classifier was trained and written by Filip Polbratt and Olav Nilsson. The training data they used was a set of 40 articles, classified as fake or real news, 20 of each type. The articles were about the presidential election 2016 and has manually been classified by C. Silverman [19]. From the training data the classifier learned which words and word combinations are used in fake or real news.

### 3.5 Analysis of the collected data

The last step is to save the data. For each URL the person who posted the tweet, the numbers of followers and retweets are saved and for each update of a Bitly block the clicks are saved. This is extracted from our collected data and put into a Excel file in order to easier analyze the data. The last part is to derive a percentage-based limit for how much of the total number of clicks a Bitly link needs to get during our collection period. This limit will give us which links will be in the final data set by filtering out links that received a large portion of the clicks before we started following it. The final data set will be analyzed to try to find similarities and differences in how followers and retweets affect the pattern of sharing of an article and if there is any difference in the collected data depending on if the article is classified as fake or real news.

### 3.6 Delimitations

**Naive Bayes classifier**

The classifier programmed the previous year by bachelor students is not very accurate, according to the thesis [18]. The classifier will not provide us with an accurate analysis of fake news, but we still think its an interesting aspect to analyze. The set of training data is not sufficient and two years old. Only looking at words and combination of words is not enough
3.6. Delimitations

to decide if an article is fake or not, facts need to be verified. Since our focus in this project is not the classifier no time will be spent improving it.

**Threshold**

Our threshold is based on a data set of 1,149 Bitly links that were followed for 12 hours and saved the number of clicks. The decision of moving links to the 24h update interval is partly based on the number of clicks the link had at hour 12. However, the bias introduced here will be eliminated by the random set of links selected and by setting the same probability for a link to be labeled random as top.

Popularity spikes are also not taken into account. We will not look at the increase of clicks from $t=0$ to $t=12$. Links moved to the 24h update interval are considered less interesting and if these links suddenly gets a large amount of clicks they will not be moved back to the 2h update interval, which may lead to loss of interesting data. Follower count or retweet count is also not taken into consideration in the threshold.

**Weekly patterns**

We will only collect tweets for 7 days which means conclusions from sharing patterns can’t be analyzed. This may include different sharing and click patterns depending on weekday or weekend, holidays or world happenings.

**Rate limits**

Bitly has the per-minute, per-hour and per-ip limits. It is also possible to expand at most 15 links on one connection, at the most 5 concurrent connections. Twitter streaming API has a 1% limit of all the tweets posted. This limit was never exceeded during our collection period, meaning all tweets containing Bitly links were collected and no data was missed.

**Update of retweets and followers**

We will not update the number of retweets and followers using the Twitter REST API at the end of each Bitly block followed for five days.

**Robustness**

Since our data mining will run for 12 days straight, the program is really sensitive to errors and crashes while running and to minimize the consequences of an error the program was made more robust. Data about Bitly links, tweets and seen Bitly links are saved to .txt files which makes it possible to restart the program if it crashes without losing any data. The data mining process was also divided into 3 separate programs, collect tweets and follow and update Bitly links, classify news articles and save data to Excel file. Everytime a successful update of a Bitly block is completed, the multiprocessing queues which is used to manage when it’s time to update a Bitly block, the queue and how many turns each Bitly block has completed is saved to a .txt file. If the program crashes or gets stuck it will lead to a short offset in the total time a Bitly block is followed, the new time will be $120\text{ hours} + \text{down time}$.

**Link shorteners**

There are multiple different link shorteners, Bitly was chosen mainly because of their statistical tools. It is also one of the most commonly used shortener. There is no statistics about specific groups only using Bitly and other groups using an other shortener, but this could be a potential bias introduced. We will not investigate if this is the case.
3.6. Delimitations

Clicks from multiple sources
It is not possible to guarantee that the clicks on the corresponding Bitly link is from the collected tweet. The Bitly API and Twitter API works independently. The Bitly API looks at the total number of clicks for that link, meaning clicks can come from multiple tweets or even other social media sources as well.

Clicks from bots
The Bitly API does not separate unique clicks on the links and multiple clicks from the same location and user. We are only interested in clicks from real humans and not from web crawlers but this is a statistical limit of the Bitly API. So it is not possible to distinguish bot clicks from human clicks. This is not a problem when looking at fake and real news, since according to paper *The spread of true and false news online* [10] the robots and bots didn’t change their main conclusions about spread speed since bots spread fake and real news at an almost equal rate. However this is a problem as we are only interested in the amount of clicks a link has obtained.
During the 7 days of collecting tweets through the Twitter API we in total gathered 10.8 million tweets containing a Bitly link. These tweets lead to 1,877,045 unique Bitly links. By using the Bitly API we got the expanded URL. By filtering on our set of news sites a total of 1,481 links of interest was obtained. By using the Bitly API we followed these for a total of 5 days. By using the classifier it was predicted if these links were classified as real or false. Approximately 25%, 366 links out of the 1,481 were classified as biased while 75%, 1115 links were by our classifier classified as real, as presented in Table 4.1. Worth noting is that the top set, random set and bottom set do not add up to 1481, and this is because as explained in Section 3 that the top set and random set have overlapping links. The categories are the ones illustrated in Figure 3.5 in Section 3. Out of the 1,481 links that were gathered 21%, 324 were retweets. When we split the links into either the 2 or 24 hour update interval 41.6%, 640 respectively 58.4%, 841 links were put into these. Our aim, as calculated in Section 3, was to move 56.25% of all links to the 24 update interval. This means that our calculations were solid with the small differences likely coming from the relatively small data set we used for our calculations. Our motivation to move this percentage was because we wanted to balance the random and top set to each 25% of all the links. In this data set the top set contained 25.5%, 378 links and the random set 24.8%, 368 links which also are numbers that are consistent with our calculations made before the collection period.
4.1 Filtration of the collected data

Because we wanted to look at mostly newly posted links and their spreading we chose to filter the total data set containing 1,481 links. This was done by excluding all links that did not receive more than a percentage of their total clicks at $t=120$ during our collection time. In order to calculate a percentage of links to remove we applied three different limits, 25%, 50% and 75%, on all links corresponding to the percentage of its total clicks the link had to receive during our collection period. The results are shown in figure 4.1. As presented, when we filter on 25% we gain a big difference, meaning we filter our links which received most of its clicks before we found the link. Relatively, the difference between 25% and a higher percentage filtration is minimal. Because we don’t want to filter out unnecessarily many tweets and the difference of the impact with larger percentages being minimal we choose to filter out links that had received at least 25% of its total amount of clicks during our collection period. We also chose to include the plots for how this filtration looks by only looking at links that were not observed until at least 24 hour into the measurement period. This was to try and minimize the number of old Bitly links that has been circulating for a time and observe if this makes a difference. The difference is minimal meaning this is no aspect worth taking into consideration because of how Bitly links are created and shared on social media. This resulted in 710 links. When looking at the sets in the filtrated data set, we get a large contrast to the numbers from the unfiltered data set. 68.5%, 486 links were in the 24 hour update interval and only 31.5%, 224 in the 2 hour interval. The top set and random set also differs a lot from the unfiltered set and therefore also from the calculations made prior to the collection period. The random set contains 163 links which is 23.0% of the filtrated set, which is reasonably close to our calculations. The top set is only 11.1% of the filtered data set, which is considerably lower than expected. This is an effect of that the tweet we collect is new but the Bitly link might be old, so we have missed its peak in popularity and gets filtered out. Many of these links has generated a lot of clicks and are therefore in our top set. Out of the filtered data set we found only 10.7%, 76 retweets and 89.3%, 634 original tweets.

![CCDF for different filtrations on the collected data at $t=120$](image)
### 4.2 Fake versus real news

Out of the filtrated data set 26.6%, 189 links were classified as fake and the remaining 73.4% as real. Because of an untrustworthy classifier we choose to only analyze the articles from BBC and Breitbart instead of all news sites. These sites were chosen because we noticed a continuous trend in the results from our classifier where BBC often was classified as real and Breitbart fake, but most importantly because these two websites are known for posting these kind of articles. This coincides with previous years thesis as well [18]. Our classifier classified 66.6% of the Breitbart articles as fake and only 12.6% of the BBC articles.

#### Table 4.1: Table with statistics of all collected data

<table>
<thead>
<tr>
<th>Sets</th>
<th>Unfiltered</th>
<th>Filtered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top set</td>
<td>378 (25.5%)</td>
<td>79 (11.1%)</td>
</tr>
<tr>
<td>Random set</td>
<td>368 (24.8%)</td>
<td>163 (23.0%)</td>
</tr>
<tr>
<td>Bottom set</td>
<td>1,103 (74.4%)</td>
<td>628 (88.7%)</td>
</tr>
</tbody>
</table>

#### Categories

<table>
<thead>
<tr>
<th>Categories</th>
<th>Unfiltered</th>
<th>Filtered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top &amp; random</td>
<td>106 (7.1%)</td>
<td>19 (2.7%)</td>
</tr>
<tr>
<td>Top &amp; not random</td>
<td>272 (18.4%)</td>
<td>60 (8.5%)</td>
</tr>
<tr>
<td>Bot &amp; random</td>
<td>262 (17.7%)</td>
<td>145 (20.4%)</td>
</tr>
<tr>
<td>Bot &amp; not random</td>
<td>841 (56.8%)</td>
<td>486 (68.4%)</td>
</tr>
<tr>
<td>All sets</td>
<td>1,481</td>
<td>710</td>
</tr>
</tbody>
</table>

#### Figure 4.2: CCDF for difference between BBC and Breitbart at t=120

Figure 4.2 shows the measurements of the links to the two chosen news sites. It is clear BBC links has a larger probability of reaching larger amounts of clicks, where BBC has a 28.79% chance of reaching over 1,000 clicks and Breitbart only has a probability of 7.14%. Worth noting is that BBC also has over a 5% chance of reaching over 10,000 clicks, while Breitbart has 0%. The equation for the trendline for BBC is $y = -0.357x + 0.4572$ while Breitbart’s is $y = -0.4801x + 0.3404$, meaning Breitbart’s CCDF is considerably more declining which also shows in the figure.
4.2. Fake versus real news

In Figure 4.3 it is observed how the other news sites in our set of news sites of interest relate to each other in terms of number of clicks. The number of collected tweets containing links to the different news sites heavily differ, with CNN only having 6 links and The Guardian having 274, which is worth taking into consideration. It is observed that links to Fox News receive considerably more clicks than the other news sites and have the biggest chance to do this.

In Table 4.2 the results conducted by Lundström et al. are presented where they used our data to calculate the medium retweet rate for the different news sites [20]. Worth noting is that the used data is only our top set, meaning the number of links will be limited. It is clear how the medium retweet rate for tweets linking to Fox News is considerably higher than for other news sites, which is also shown when looking at all links in Figure 4.3. It is also observed how many more tweets contains links to Breitbart than BBC in our top set but tweets of BBC still being retweeted more frequently, coinciding with the result presented in Figure 4.2.

Table 4.2: Medium retweet rate for different news sites for our top set [20]

<table>
<thead>
<tr>
<th>News Site</th>
<th>Medium Retweet Rate</th>
<th>Number of Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fox News</td>
<td>7.6</td>
<td>10</td>
</tr>
<tr>
<td>The Times</td>
<td>2.0</td>
<td>1</td>
</tr>
<tr>
<td>BBC</td>
<td>2.0</td>
<td>24</td>
</tr>
<tr>
<td>CNN</td>
<td>0.5</td>
<td>2</td>
</tr>
<tr>
<td>Breitbart</td>
<td>0.4</td>
<td>52</td>
</tr>
<tr>
<td>The Guardian</td>
<td>0.2</td>
<td>74</td>
</tr>
<tr>
<td>Huffington Post</td>
<td>0.2</td>
<td>23</td>
</tr>
</tbody>
</table>
4.3 Difference between categories and sets

When looking at the different sets of links: top, random and bottom we get the four different categories of links presented in Chapter 3, Figure 3.5. The results of each sets’ clicks at $t=120$ for these are presented in Figure 4.4. One thing worth noting is the similarity between the top set and the top and random set with only a couple of links in the top set that obtained more clicks. This means that the links in the top set that was not labeled as random follow the same pattern as those labeled as random. The link that got most clicks had 2,028 clicks at $t=0$ and 20,524 at $t=120$, an increase of 1012%. The article was posted on April 14th, three days after we started our collection. The title of the article was "US strikes Syria after suspected chemical attack by Assad regime" and the subject is both popular and controversial which might be a reason for its spike in popularity.

![Figure 4.4: CCDF for all data sets at $t=120$](image1)

![Figure 4.5: CCDF for the random set versus all sets at $t=120$](image2)
4.3. Difference between categories and sets

In Figure 4.5 the similarity between all sets and the random set is easily observed, meaning that the random set is a good representation of all links. This is both desired and expected because the likelihood of a link being labeled as random being the same for links above the threshold as below. The equation of the trendline for all sets is $y = -0.4209x + 0.2399$ and for the random set it is $y = -0.3996x + 0.1742$, which shows the similarity.

![Figure 4.6: CCDF at different timestamps for top, random and all sets](image)

By instead looking at the clicks a link has obtained at different timestamps a different result is received. The result of looking only at the top set is presented in Figure 4.6a. The result of looking only at the random set is presented in Figure 4.6b and all sets are presented in Figure 4.6c. It is observed how the CCDFs for the top set longer stays closer to 1 in relation to the random set. The link with the least number of clicks in the top set at $t=120$ has 473, which means that it according to our threshold at 450 received only 23 clicks from $t=12$ to $t=120$. All CCDFs also end at a number of clicks relatively close for all timestamps, spanning from 17,245 for $t=2$ to 20,524 for $t=120$. This whole observation is caused by a single link rapidly getting clicks in the first two hours and barely getting over the 25% filtering limit. What is also worth noting is that these number of clicks are not from the same link. The one with 17,245 at $t=2$ received 14,876 clicks in the first 2 hours of our collection period but 2,015 clicks for the remaining 118 hours.

The random data set’s CCDFs dips a lot sooner than the top set. This is based on that links with a small amount of clicks may be labeled as random. Another comparison between the random data set and all sets at different timestamps show the resembling appearance, which also was shown in Figure 4.5. Looking at $y = 10^{-1}$, meaning it has a probability of 10% to obtain a number of clicks for the different timestamps: for $t=2$ it has a 10% probability of
4.3. Difference between categories and sets

reaching 347 clicks, for \( t=4 \) it was 585 clicks, for \( t=8 \) it was 755 clicks, for \( t=12 \) it was 832 clicks, for \( t=24 \) it was 927 clicks and for \( t=120 \) it was 1,046 clicks.

Figure 4.7: Clicks at \( t=12 \) versus \( t=120 \)

Figure 4.8: Clicks at \( t=24 \) versus \( t=120 \)

If we follow the \( y=x \) lines in Figure 4.8 and 4.7 we can see that that the plots look very similar, both have dots close to the line which means that they receive close to zero clicks between the 12 or 24 hour mark to \( t=120 \). The big difference is that in Figure 4.8 the dots are more scattered above the line. This support the theory presented in Chapter 2.1 that a tweets activity is essentially concentrated around the first hours of its existence.
4.3. Difference between categories and sets

We can in Figure 4.9a see that the majority of links in the top set, which was links that had more than 450 clicks at \( t=12 \), continue gaining clicks after \( t=24 \). The average link obtains 340 clicks from \( t=24 \) to \( t=120 \), which is a small number if compared to the largest gain of 3,865 clicks. The link that received the smallest amount of clicks received 5 clicks. If we instead look at the random set presented in Figure 4.9b we get an average of added clicks as 62. The link that received the most number of clicks in this time period received 3,014. In this set we have several links that did not receive any clicks from \( t=24 \) to \( t=120 \), which is a result of labeling links as random regardless of their clicks, which is an indication that the threshold we introduced is good guideline for knowing if a link is worth updating more frequently. As shown earlier, the random set is also here a good representation of all links shown by the similarity in Figure 4.9b and 4.9c.
The Pearson correlation of all sets in Figure 4.9c is 0.46, which represents how good the correlation between clicks at \( t=24 \) and clicks added is. The equation for the trendline is \( y = 0.0183x + 1.210 \), which means that every \( 1/0.0183 = 54.64 \) clicks a link has at \( t=24 \), it will receive one additional click up until \( t=120 \). The lower and upper bounds in the figure illustrate a 95\% confidence interval. The average number of added clicks for all sets is 64.

It is observed how the placement of dots in Figure 4.10a is substantially less diversified than in Figure 4.10b. This is expected because of how the random set can contain of links that do not receive a single click between \( t=2 \) and \( t=4 \), shown by the dots located on the x-axis. The two links that received the largest number of clicks from in this interval had 1,134 and 9,279 at \( t=2 \) and received an additional 585 respectively 583 clicks until \( t=4 \). This is an increase of 51.5\% respectively 6\% in number of clicks. For the top set the two links that received the largest number of clicks had 2,466 and 4,809 clicks at \( t=2 \) and received an additional 1,247 respectively 1514 clicks until \( t=4 \). This imply an increase of 50.5\% respectively 31.4\%. Even
in the top set there were links that received 0 clicks from $t=2$ to $t=4$. The Pearson correlation of all sets in Figure 4.10c is 0.61, compared to the 0.46 of all sets in the interval $t=24$ to $t=120$ in Figure 4.9c. The difference in correlation could come from the difference in time span, where there is a stronger correlation when looking at the two hour interval compared to the 96 hour interval. The equation for the trendline is $y = 2.401x - 1.637$, which mean that every $1/2.401 = 0.4164$ clicks a link has at $t=2$, it will receive one additional click up until $t=4$. This largely differs from the trendline in Figure 4.9c where it was 54.64, which is caused by the large amount of clicks a link gets relative to the difference in the length of time periods. The lower and upper bounds in the figure illustrate a 95% confidence interval.

By looking at the three figures in Figure 4.10 and comparing these with the three in Figure 4.9 we can note several key observations. All the different sets has overall similar appearance for both the time intervals. The results for the top set in Figure 4.10a has as expected less clicks at the start time than in Figure 4.9a but has a similar appearance in the clicks added. The same differences and similarities apply for the two plots for the random sets and for all sets as well. This mean that the relative addition of clicks in the interval from $t=2$ to $t=4$ is similar to the addition from $t=24$ to $t=120$. This implies that our threshold at $t=12$ could be moved to a more prior time in the collection period, and would possibly yield a similar result. Looking at our figures it may even be conceivable to put the threshold at $t=4$.

![Figure 4.11: Pearson correlation $r$ of clicks at t and t+T](image)

In Figure 4.11 we can see the Pearson correlation between clicks at time $t$ and time $t+T$, where $t$ is the sample point and $T$ is how far forward we look. We can see that the curve representing $t=0$ has the lowest Pearson values $r$, which is expected since it’s harder to predict a links future clicks earlier in its existence. The curve representing $t=12$, which was the time where we implemented our threshold, has the highest Pearson values $r$, which indicates that having our threshold at $t=12$ was a good choice.
4.4 Impact from a user’s of number of followers

In Figure 4.12 we can observe how the number of followers of a tweeter impact the number of clicks a link posted by the user gets. A tweeter with more than one million followers has a guaranteed number of 245 clicks and has at least a 10% chance to reach 10,000 clicks. Tweeters with less than 1,000 followers can according to our data set obtain a higher total amount of clicks than the ones with more than a million followers, but only have a 1.88% chance of doing this. Worth noting is that tweeters with less than 1,000 followers have both a chance to receive a larger number of clicks and a larger probability to do so than those with followers between 1,000 and 100,000, which seem unreasonable. Worth noting is that the 2 links that obtained the largest amount of clicks in our data set were linked by a user with less than thousand followers, meaning that there is no strong correlation between number of followers and clicks, as discussed in Section 2.1. This may be a result of our relatively small dataset compared to other studies in the field, but may also prove that spreading of news on social media may have a random factor to it. When adding a Bitly link to a tweet, the only information about what the link leads to is what the tweeter writes about it in the tweet. Therefore the potential random factor can be how good the tweeter is at generating an interest for the potential clicker. This means that popularity and influence is affected by more factors than just followers. We won’t look at the users that posted the tweet or investigate if this lead to more followers because of the ethical reasons explained in Section 5.3.
In Figure 4.13 we can see how the number of followers influence the number of clicks a link to an article receives. If we follow the line $y = x$ we can see the ratio clicks/followers is less than 1. The majority of the links has less than 1 added click per follower, which is shown by all the dots being bellow the orange line and the Pearson correlation being 0.00094. This means that a larger number of followers is correlated, but not strongly, with obtaining more clicks which support the result presented in Figure 4.12.
In the following chapter we will discuss our results and implementation of our methodology. We will also discuss how our methodology can be used for other studies and our work in wider context.

5.1 Results

When looking at how the number of followers influence the number of clicks on a link, we get a result that coincide with the results Cha et al. received in their study as presented in Chapter 2. They presented that being a popular tweeter with a larger follower base is not strongly correlated with obtaining more clicks [5]. This indicates that there is several more factors than just the follower base which impacts on how news spread on social media.

One thing that did not coincide with the results discussed in 2.2 was that BBC had a larger probability of reaching a larger amount of clicks, shown in Figure 4. Vosoughi et al. came to the conclusion that falsehood spread content faster, farther and deeper in all categories of news. The difference in our results may depend on our limited data of links to BBC and Breitbart as well as BBC simply being a more popular website with more visitors, which is something we are not taking into consideration in the results.

It is in the results also observed that the random set is a good representation of the set with all links. This is expected because of how we label a link as random regardless of its number of clicks. This also means that the random set consists of less links and this would mean that some aspects of our results would be relatively resembling with a smaller data set. This indicate that, even with our limited data set, the characteristics of spreading of news on social media presented in our results is scalable to a bigger scope. It is also observed that a link can generate considerable amounts of clicks several days after it is posted, results that correspond with the results Gabielkov et al. discovered in their study as mentioned in Chapter 2.1 [6].
As observed and discussed in the results there are comprehensive similarities in growth of clicks from $t=2$ to $t=4$ and $t=24$ to $t=120$, which indicate that the threshold could be set at an earlier time than $t=12$. By looking in our top set at the link with the least amount of clicks at each timestamp we get that the optimal threshold - if you would choose to put the threshold earlier - should be put at $t=6$. This would mean that 178 links were put in the top set which would balance the top and random set in the filtrated data set. This is a massive increase in the size of our top set but by looking at Figure 5.1 it would be the best choice. This is because by choosing $t=6$, $t=8$, or $t=10$ would approximately give the same result, but by setting the threshold at $t=4$ we would more than quadruple the top set, which is undesired.

This would be of importance for this thesis project’s main focus, reducing the number of Bitly calls needed in order to collect more data. This is because more links could be put into the 24 hour update interval earlier in the collection period and therefore update less number of times in total. This implies a high scalability for this project.

5.2 Method

When looking at our two data sets, the unfiltered set with all links and the filtrated one with links that received at least 25% of a link’s total clicks during our collection period, we get a substantial difference in their accuracy in relation to our set calculations made prior to the collection period in order to find a suitable threshold. This is because of two things.

Firstly, a lot of the links that got filtrated out had extensive amounts of clicks when before we found them. Most notably the difference was in the top set which went from 25.5% to 11.2%. This somewhat expected, but not in this great extend. The explanation to why our calculations did not fit our filtrated data was because the calculations were made on all links without the filtration limit. Instead we needed to filter the data set made prior to the collection period to find a suitable threshold in the same way we made on the collected data. This would suggest that in our unfiltered data set the random set would be 18% and the top set 32% of all links in order to balance the top and random set after the filtration.

Secondly, our data set which was used to derive the click threshold, was only followed for 12 hours and instead we should have followed and updated the links for 5 days, as we do in our study. With our three filtration limits, 25%, 50% and 75%, this would have taken a total of 7 days. This was not manageable in our timeline.

We could have also chosen a different methodology in how we chose if a link should be moved to the 2 or 24 hour update interval. By instead looking at the number of clicks it re-
ceived from $t=0$ to the threshold you could filter out both links that would not get many clicks at all during the collection period, but also identify links that received a lot of clicks during this time and therefore possibly have their largest peak in popularity. This could mean that we would only find links that already are popular and in that way miss articles that will reach their peak later than in this interval, which would be a drawback of this implementation. In our limited time span we had to fully implement our methodology and only afterwards collect data. The program was a continuation of an already built data mining tool. A possible consequence of building on a program made for a different study and the limited time span is how effectively the program collected and processed tweets and links. This could lead to a limited data set, which is a drawback. In order to fix this we discussed with the people who developed the basis and tried to figure out solutions to the errors that occurred on unpredictable conditions during our collection period.

We also used a classifier not written by us, which may have lead to inaccurate results when looking at the aspect of real or fake news. This was not evaluated because it was not the focus of this thesis project.

5.3 The work in a wider context

The data set collected for this thesis contains tweets and Bitly links with clicks at different timestamps. As stated by Gabielkov et al. a general rule is that data sets containing clicks should be considered as highly sensitive information [6]. Since we use the Twitter streaming API we will only get tweets that are categorized as public in our tweet data set. This data set can potentially be harmful to individuals since it can be used for targeted advertisement. Our Bitly data set only contains "short URL" and "long URL" to the news article, the number of followers the tweeter has and the tweet’s number of retweets. We do not save anything that directly links an individual posting the tweet to the link. The number of clicks in combination with the URL can on the other hand be of great value to companies. Most importantly none of our data sets will be made publicly available.
Our thesis is heavily focused around our methodology. We have created a working methodology, with the help of our threshold, to help decide which Bitly links are interesting and which are not. With this implementation others are able to, in a future data collection, only collect and follow Bitly links that are interesting and therefore reduce the number of calls to the Bitly API. This would enable an increase in the scalability which would result in a larger collection of data that only contains interesting links to analyze. This gave us a great data set in order to examine the popularity dynamics of news articles and their spreading on social media. Since our classifier, used to classify articles as fake or real is not trustworthy, our work can be seen as groundwork for a better methodology regarding classification. It is observed that there is no strong correlation between number of followers and number of clicks, indicating that there are several factors that affect how news spread on social media. Our random set is a good representation of our complete data set, meaning a smaller data set would derive similar results. Secondly, it is observed that the spreading of news on social media follow a pattern. Most of a links clicks are obtained in the first hours of its existence and level out afterwards. At last, a threshold can to an extent determine if a link is interesting and worth following more frequently. Most of the links with the largest peaks in clicks were placed in the top set. This threshold can be set at different times and our result indicate that it would be convenient, in consideration to the number of Bitly calls, to set the threshold earlier than $t=12$.

6.1 Future work

Further development and experimentation in our way to decide which links that are interesting is definitely possible. Instead of using a constant like we did, where a link that has less clicks than 450 at $t=12$ is considered not interesting, there are several other possibilities. For example using a percentage-based threshold would be interesting to investigate, where a link is considered interesting if it from $t=0$ to $t=12$ gains a percentage more clicks. This percentage-based implementation would probably work better for filtrating out links with a large amount of clicks when we found them, these links have probably been circulating for a long period of time.
6.1. Future work

It would be interesting to see what observations and conclusions our data can give with a better classifier. A different classification methodology is to use fact-checking organizations such as snopes.com, politifact.com, factcheck.org, truthorfiction.com, hoax-slayer.com and urbanlegends.about.com. These sites doesn’t have any APIs but it’s possible to parse the title, body and verdict to create a large set to compare the articles found in the collected data set with.
Bibliography


