# PROBABILITY OF SENDING MALICIOUS CONTENT TO THE FRONT PAGE OF REDDIT PREDICTION

**Andrew Fong A09479966**

**Shirley Quach A09491420**

**CSE 190: Data Mining**

## Introduction

Reddit is a message board whose content is fueled by posts made by users that are "upvoted" and "downvoted" by the rest of the community to determine the post's position on the site. Popular content will end up in the front page and unpopular content will most likely never see the light. Reddit allows the ability to comment on posts and encourage discussion to determine its legitimacy or for a user to simply add a relevant anecdote pertaining to the post. The front page of Reddit consists of posts from various communities, called "subreddits".

With large internet communities such as Reddit, there is a lot of traffic coming from various parts of the world. We believe he front page of Reddit acts as a potential breeding ground for malicious attacks.

For a small group of malicious users, their goal is to spread malicious intent to as many users as possible. This can be done by taking advantage of previously popular content and getting voted to the top where the content will seem worthy of opening. We want to predict what types of communities are susceptible to these attacks.

## Exploratory Analysis

In this assignment, we are analyzing a Reddit submissions dataset. The dataset is a collection of 132,308 image submissions containing 16,736 unique image submissions. This dataset looks at resubmissions of an image and collects features such as its rating, the title, and the number of comments received. The information is captured in one file dating from July 2008 to January 2013 containing: image id, unix time, raw time, title, total votes, Reddit ID, number of up-votes, subreddit, number of down-votes, local time, score, number of comments, and username.

This data was originally used to optimize reposts by determining the best post title, time of post, community posted to and type of content in order to maximize the number of upvotes for the posting user.

Table 1 includes initial data obtained from the dataset in addition to some data we thought was relevant for us to know.

We thought finding the number of post titles with the word "xpost", "crosspost" or cross post" to be important. A title with any of the words listed indicate that user had taken the post form another subreddit and wanted to

share that post to another subreddit in which other users may not have seen yet and is relevant to the community. We thought a user posting with those words in the title would make another user less likely to downvote the content as the original poster did not claim the content as their own and a source was listed.

**Table 1: Dataset Findings**

| | |
|---|---|
| # Unique Reported Users | 57, 341 – 58,474 |
| # Unique Images | 16,670 – 16,736 |
| Total Submissions | 117,007 – 132,308 |
| Avg Repost | 7.019 – 7.9 |
| Avg Repost/Person | 1.59 – 1.73 |
| Submissions with Verifiable Usernames | 99,198 – 114,498 |
| # of Crossposts (in Training Data) | 678<br><br>(292) |
| Avg score of Crossposts (Training Data) | 263.072<br><br>(247.39) |
| Avg score of all posts (Training Data) | 231.088<br><br>(242.877) |
| Most Posts | Gangsta_Raper, 5,217 |

In Figure 1, the data indicates the average number of reposts person. The original data contains outliers like a user having posted 5,217 posts on a single account. When calculating averages of anything by a user, it is important to take the outlier into consideration.

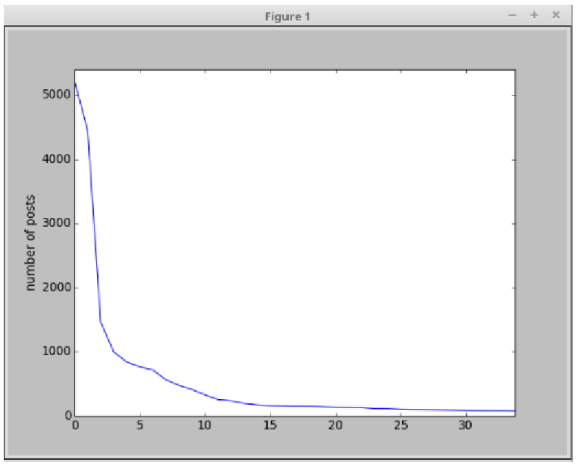


Figure 1: Number of Posts per User

## Predictive Tasks

In this assignment, we are interested in determining whether or not an image post with malicious intent can be effective by simply being on the front page of Reddit and if it has the ability to continute to propagate. If it has reached the front page of a given subreddit, it implies that it has reached maximum visibility.

We want to determine the probability of a repost to be in the front page of Reddit or any give subreddit. If an attack can reach the front page and receive full visibility then we will consider it a successful attempt where many unsuspecting users may click on the image for any number of reasons.

We decided to use a predictive model in order to determine whether a post will be high enough to be in the front page where malicious content will hopefully propagate even more. In the study by Lakkaraju, McAuley and Leskovec on popular front page posts it indicates that there is a relationship where the top posts will receive more views which will lead to even more views.

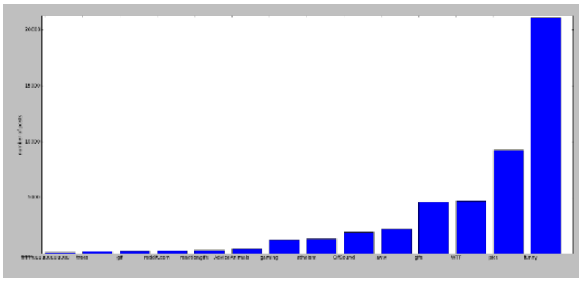

Figure 2: 14 Most Popular Subreddits to Make a Repost

In Figure 2, we found the the top subreddits to make a repost in include Funny, Pics, WTF, GIFS, GIFsound, Atheism, Gaming and Advice Animals.

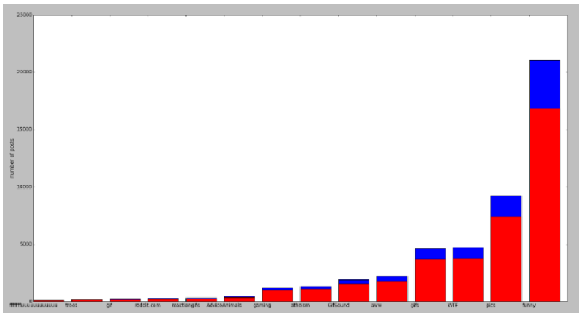

Figure 3: Section we are Aiming for

In Figure 3, the blue area indicates the top 20% of a subreddit malicious content will aim for in order to achieve maximum number of views possible.

## Related Literature

Literature we thought was relevant was a study on Reddit titled “What’s in a name? Understanding the Interplay between Titles, Content, and Communities in Social Media” by Himabindu Lakkaraju, Julian McAuley and Jure Leskovec. This study looks to find the best post title, time, subreddit and optimal time for a post to be created and upvoted to the front page.

This study uses a community model which looks at how a subreddit responds to submissions. A language model which takes a look at the type of language used to for a submission. Some state of the art technique used to study this dataset includes applying a weighted Jaccard similarity of a current title.

The most useful information we found from this study was the best time to make a post and learning about how some subreddits are less welcoming of posts unrelated to the community topic such as the subreddits titled Atheists and Gaming.

Another source of literature we found to be very interesting was an article by Swati Khandelwal titled “How to Hack a Computer Using Just an Image.” Khandelwal describes a new technique discovered by security researcher Saumil Shah dubbed “Stegosploit” in which it allows a hacker to upload hidden malicious code inside the pixels of an image making it impossible for one to spot when a user opens it.

Stegosploit is an obfuscation method utilizing a polyglot jpg/javascript data format to hide the fact that an image is actually executable code. While untested on popular image sharing sites, such as dropbox or imgur, the exploit has been (for a very long time) a problem that can occur on any untrusted domain. However, this report has brought to light that this is still a very relevant issue. We delved into this

project given the possibility that these types of attacks become more prevalent. There are a lot of variables that could limit any patterns from coming to light. Therefore, we first had to make a set of assumptions in order to focus on the threat we are analyzing.

Similar to Reddit and our goal is another study titled “Predicting Responses to Microblog Posts” by Yoav Artzi, Michael Gamon and Patrick Pantel in which they look at another social media site called Twitter in which they look into predicting responses and retweets for various posts ranging in popularity. This is similar to what we are attempting to study because it looks what posts are more likely to elicit a response in the form of retweet and comments which is the goal of a malicious image.

## Results and Conclusions

We used a logistic regression to classify whether or not an entry would make it into the front page of Reddit or a subreddit. To model this data we took the top 40% most popular entries of all the available data from the dataset and assumed that would be what an attacker would use. The initial baseline model that we used was that we took the top 20% of all subreddits and predicted that those would make it into the front page of Reddit. The datasets were not evenly distributed between subreddits as it tended to be biased towards the more popular and larger subreddits with more user views. We included two biases based off of the mean popularity rating of an image and based off the relative size of the subreddit.

From our results, we have concluded that it would be reasonable to assume that more popular images are more likely to make it to the top assuming the user is taking the more popular posts and randomly posting it to various subreddits. We have assumed that the user is very likely to automate this task and will not take into consideration the type of subreddits posted to which may decrease its chances to reach the top. In contrast, the subreddit bias that we did not obtain, it did not seem to have any clear affect and clear patterns weren’t able to be distinguished. We assumed larger subreddits tend to be eclipse smaller subreddits and smaller subreddits tends to be supersets of larger subreddits. This implies, users who would in these smaller subreddits may have already witnessed the post at a previous time.

In our attempts to differentiate out methods from the similar studies we used, we disregarded a good amount of variables.

## Challenges

We wanted to look at how the comment section for a post can be just as effective for an attacker to do damage. From personal experience, we noticed that many highly upvoted comments are from users posting links to images in order to express their sarcasm or feelings more effectively which will faster elicit a reaction from the reader as opposed to reading a wall of text. We believe the comments section may be a contributing factor and have some influence on whether a post gets upvoted to the top as well. We hypothesized that positive responses and contributions may be a factor in pushing a post to the top and posts receiving negative responses such as indicating the post was a repost will encourage downvotes. If one user recognizes the post there is a good chance that thousands of other users will recognize

the post as a repost as well. Unfortunately, the dataset we used did not take comments into consideration.

We realized it would be another challenge to analyze text data, but another idea would be to look at whether to look at whether a comment was given Reddit gold also known as ‘gilding’ a comment. Reddit gold is a premium membership program that comes with extra features on the site.

Another user is able to give Reddit gold directly to a user or gilding one of their posts/comments. Single gilding costs $3.99 which to some is not a trivial amount.

We hypothesized the top comments worthy of being gilded may indicate the post was worth reading and a positive contribution to the community in turn will encourage upvotes which would be something interesting to look into in the future.