NTDS Project
Identifying Spammers on Social Networks

Görkem Camli, Murat Genc, Ilija Gjorgjiev, Raphael Laporte

I. Introduction

Social Networks are an essential part of many people's daily life. The growth and spread of social networks lead to an increase in the amount of spammers with bad intentions. In this project, we aim to identify the characteristics of a spammer and whether we can find them based on the features such as age, gender and time of interaction. After mastering the nature of our network data, our ultimate goal is to predict the spammers on social networks with high accuracy.

In this project we use the 'Spammers on Social Network' data which includes users and relations information in a given social network. The user features and the links within the network are essential to answer the questions of "How can we identify spammers in social networks based on their profiles?" and "Can we group the spammers based on their features?".

II. Data

The Spammers on Social Network data (1) consists of users and relations. For both the network analysis part and classifier part, due to the heavy computation and execution constraints, we used a sub-sampled version of the data.

A. Data Preparation for Network

The preparation of the data consists of defining nodes as users, edges as relations and creating the corresponding adjacency matrix. The network created is filtered on the relation 5, number of links in the network is 83176 and average degree is 1.33.

In order to have a better visualization of this network, a graph is created and manipulated by using a tool, Gephi. The figure below illustrates the network of spammers and non-spammers, colored in green and red respectively.

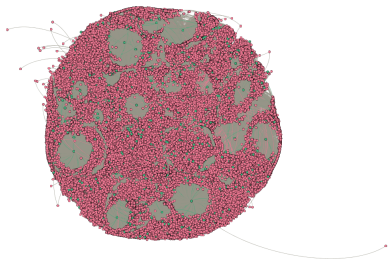

Fig. 1. Network of Spammers and Non-Spammers

III. Network Discovery and Analysis

In Network Discovery part, we calculated the basic statistics about network such as number of links, average degree, degree distribution and number of components in the network. According to the results, our network is connected; thus it has one big component. According to its degree distribution and average degree of 1.33, our network is scarce and we have few nodes that have high number of incoming and outgoing edges.

As a further step, we looked for an answer for some questions related to node and edge features that might be insightful to better understand the network and data.

How many of the users are manually labeled as spammers?

From the Table I, we can see that our data is highly unbalanced. The Spammers only around 4% of the whole network nodes. In total we have 2448 of the users identified manually as spammers.

Spammer Label	Count
Non-Spammer (0)	59725
Spammer (1)	2448

TABLE I
Spammer Distribution for the Users

In real word, this number is actually very serious, since this means one out of every 25 people is spammer. This shows us again that detecting spammers are essential and important task to achieve.

What are the age distributions for spammers and non-spammers?

Age is one of the user features that we thought might be highly correlated with the spammer detection. Even by only looking at the network data, we can see from the Figure 2, most of the users actually fall in the age range of 20-50.

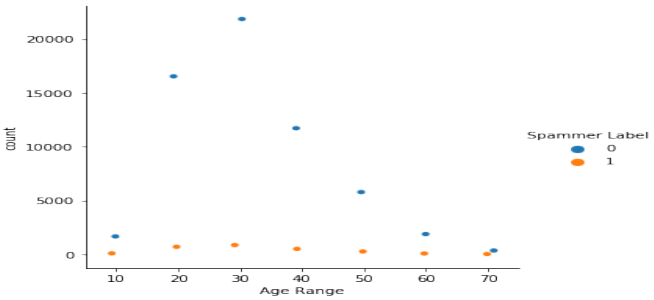

Fig. 2. Age Distribution for Spammers and Non-Spammers

When we look deeper in the spammer age distribution only, we can see that the age pattern still exists and the majority of the spammers are within the age range of 20-50. Therefore, we can actually conclude that this feature might have an important role on detecting spammers.

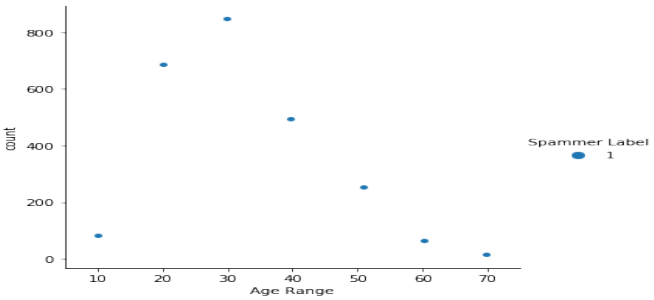

Fig. 3. Spammer Age Distribution

What are the gender distributions for spammers and non-spammers?

Gender is also another important aspect that can be thought of while identifying the spammers. Before the prediction, we checked how gender balance is distributed in the data and overall in the spammer group.

Spammer Label	Gender	count
Non-Spammer (0)	F	77
Non-Spammer (0)	M	59648
Spammer (1)	F	39
Spammer (1)	M	2409

TABLE II
GENDER DISTRIBUTION FOR THE USERS

From the Table II, we can see that there is no gender balance in the data. Within the spammers, only 1.593% of them are females. This feature might be also distinctive while finding the Spammers.

What is the time spent by spammers and non-spammers?

Once we look through the overall time spent with all users, we can't really identify spammers within the given time range from Figure 4. However, if we look specifically for the time spent for spammers, we actually see that most of the spammers are accumulated on the lower part of the plot suggesting that they do their activities in an interval of less than 0.2. This actually makes sense when we consider that spammers are generally spam people who do short events too many times; such as sending pokes, mass emails etc.

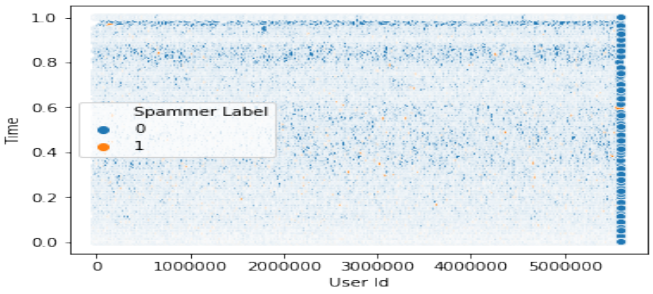

Fig. 4. Time Spent Distribution

One of the interesting facts that we have found is that non-spammers are actually spending more time on social network. Although the overall comparison of total time spent by spammers and non-spammers on social network doesn't give us a direct information, it might be useful to focus on the time spent distribution plot of the spammers. By this way, we can actually relate some of the regular activities which spammers pursue.

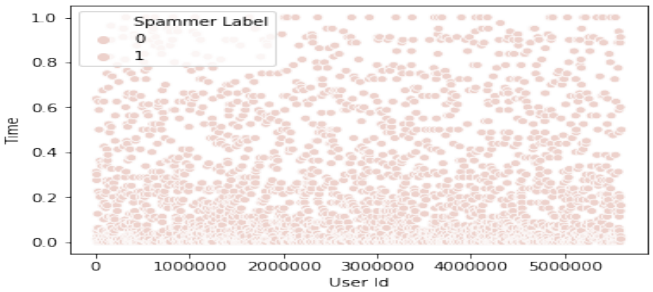

Fig. 5. Time Spent Distribution for Spammers

Therefore, time spent might not be as distinctive as gender or age distribution while identifying; but certainly gives us some important intuition that we can relate spammer activities in real life.

How many relations belong to spammers? What % of the relations they are?

Why do we call spammers as spammers? Oxford Dictionary defines spammers as "A person or organization that sends irrelevant or unsolicited messages over the Internet, typically to large numbers of users, for the purposes of advertising, phishing, spreading malware, etc". With this idea in mind, we want to check how many of the relations (activities) are

actually started by the spammers in our data. According to our results, 62055 out of 83176 edges starting source belongs to Spammers. This is around 74.6% of the total edges in our data set. This number is very high, especially when we consider that only around 4% of the all users were spammers. This means that very few users account for almost 75% of our network. By referring to this information, we can actually understand our network better by concluding that majority of our dense nodes could be the spammers.

What are the number of different days spammers interact? (same for non-spammers)

Another information we check also how many different days the spammers performed their activities. Normally, spammers spam people in a short time interval with huge amount of activities.

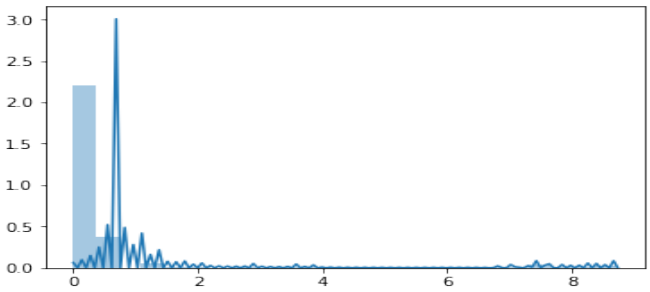

Fig. 6. Different Day Numbers Distribution for Spammers

From Figure 6, we can see that our thought holds and most of the spammers perform their Spamming activities either in 1 or 2 days. Doing a huge amount of activity in a given limited time frame can also be considered as an important property of being a spammer.

IV. Classifier and Spammer Prediction

To label spammers, we chose to use artificial neural networks. Because of how the success of this type of analysis is particularly dependant on the quality and breadth of the data that is being fed to the network, we had an interest in leveraging the size of our dataset.

A. Implications of exploring the complete dataset

The complete dataset contains 850 million relations and 5 million users. Such a size means that both the adjacency matrices and the set of relations were both too large to have either reside in memory.

If we wanted to take advantage of the breadth of the dataset, we quickly realized that we could only study the local neighborhood of each node. Such a graph size is prohibitive of any analysis requiring knowledge of the whole graph, like graph spectral analysis for example.

As such, we made an attempt to extract the local neighborhood around each node in the complete dataset. To do this, we first broke up the relationships by type, then explored the local neighborhood or each node for a given type. Unfortunately, this method requires us to store all relations of a given type in memory, which was not possible for relations of type 5 and 6.

This then forced us to subsample the studied nodes. Since the filtered nodes given to us at the beginning of the semester were known to form a connected component, we chose to use these filtered nodes, as we expected to the higher connectivity of the graph at these points. It was at this point that we realized that the sub-sampled graph that was given to us was nothing other than a 2-hop neighborhood around a particularly prolific spammer.

B. Our custom filtering

The way in which the graph was initially filtered did introduce a bias into the network : nodes that were at a distance of 2 from the central spammer node were extremely likely to have most of their relations discarded - this artificially distorted their topology, which was bound to show up in any features we extracted. Acordingly, all of our networks using features from the initially filtered dataset performed poorly.

We adopted two different sampling strategies in order to address this:

- 1. Keep all nodes, but only look at its out-degrees and in-degrees per feature type.
- 2. Choose a subset of nodes, and look at relationships of type 5 within that set of nodes. Drop any nodes without any relationships. This left us with a set of 160 thousand nodes.

With each of these two samples, we ran different networks in order to compare the advantages of keeping the overarching network structure versus only looking at local features.

C. Some of our Classifiers

Both neural networks used a series of dense layers, with an F1-score loss function. F1-score was chosen here because of the high ratio between spammers and non-spammers.

- The classifier used on the first sampling used out-degrees and in-degrees per feature relation type, sex, age and temporal information as features. Four dense layers were used here. This classifier was successful : 56-58% accuracy for spammers and 96% accuracy for non-spammers.
- The classifier used on the second sampling used the columns of the adjacency matrix and of the normalized Laplacian projected onto the eigenvectors from our sample's spectral decomposition. Two dense layers were used here. This approach was unsuccessful, resulting in at most 20% accuracy for spammers and a highly variable accuracy for non-spammers. We hypothesize that this could be because of:
  - The Sampling Method: looking at the relationships between 3.2% of nodes could be too sparse of a sampling in order to capture the overarching network structure.
  - Not computing enough eigenvectors: to remain at reasonable computation times we choose to compute

only the 100 eigenvectors of the graph corresponding to the largest eigenvalues. Even if the largest eigenvectors represent the dimensions along which the graph carries the most entropy, using 100 eigenvectors might not be enough .
- Destroying the difference between incoming and outgoing edges in order to keep the Normalized Laplacian hermitian.

D. Successful Classifier

We work with the whole dataset. Basically we read the original user dataset and save it. After that we start reading the original relations dataset line by line from the "relations.csv" file and we create 2 dictionaries.

One of the dictionaries contains the sum of each relation type with respect to each node, where the node was the source, and the other dictionary contains the sum of each relation type with respect to each node, where the node was the destination. In simple words for each node and for each relation type of that node, we get the in-degree and out-degree sum.

V. Conclusion and Future Work

In conclusion during in this project, we did further analysis on the data and network to find more insights about the data. We did some sucesful and unsucessful approaches to predict the spammers. Then do the classification to predict the spammers in social networks.

First of all for getting a better classification, we would definitely need more users that are spammers. The dataset that we worked with is pretty skewed as it contains around 7% spammers and 93% non-spammers, which makes it not an easy classification problem. We have added additional 14 features to the original ones and this gave us a 56-58% classification of spammers and 96% classification of non-spammers on the test set. For getting a better F1 score overall and better spammer and non-spammer accuracy we would need to dive deeper relations. We never worked with the timestamp that was contained in the relations data frame. After all the work and attempts that were done at classifying, we think that maybe the most important factor would be exactly the timestamps. There are a lot of real-life and logical examples which we could connect to timestamps like:

- Making an abnormal amount of relations of the same type or different types in a short period, would probably mark you as a spammer.
- PageRank of Each Node.
- Connecting the TimePassedValidation feature to timestamps of relations.
- Diving into whether the user is only doing relations and not getting any response to the same relations.

References

[1] S. Fakhraei, J. Foulds, M. Shashanka, and L. Getoor, "Collective spammer detection in evolving multi-relational social networks," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 1769–1778, ACM, 2015.

[2] N. Donges, "How to build a neural network with keras."

[3] "Best loss function for f1-score metric."