A Network Tour of Data Science

Identifying Spammers on Social Networks

Görkem Camli, Murat Genc, Ilija Gjorgjiev, Raphael Laporte

École Polytechnique Fédérale de Lausanne

23 January 2019

Outline

Introduction

Problem Description

Our project aims to analyze a user network to identify spammers and non-spammers.

Challenge

The main challenge is that there is no concrete method to identify spammers on social networks just by referring to their user features. The accuracy rate of previously published models are not high, so it has been an issue in the last part of our project.

Goal

Our goal is to interpret and understand the dataset and create a model to make predictions to determine whether a user is a spammer or not.

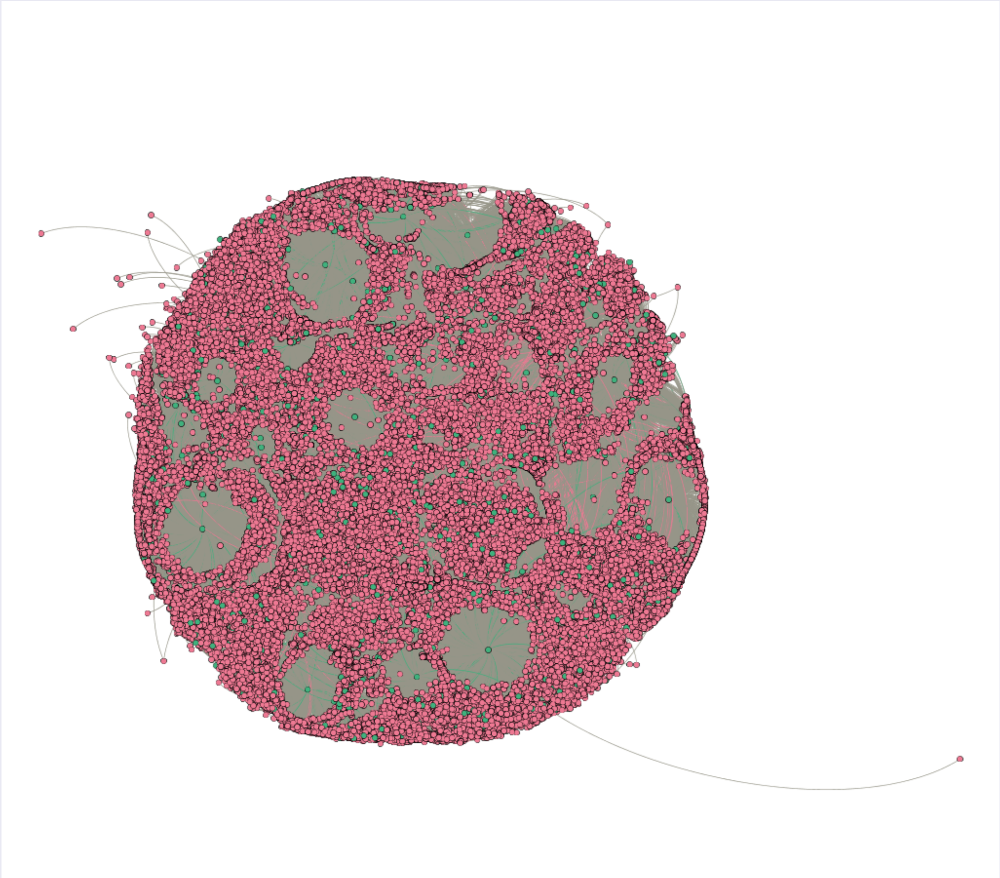
Project Goals

During this work we looked for an answer for the following questions:

1. How can identify spammers on social networks based on their profiles?
2. Can we group spammers based on their futures?

Dataset

The Spammers on Social Network Data



Basic Network Statistics

Network Properties

- 62173 nodes, 83176 edges
- Connected (one big component)
- Average Degree: 1.33
- Max Degree: 8216
- Sparse with few nodes with many in-coming and out-going edges

As a next step, to get more insight from the network and the data, we tried to answer a few questions before doing the prediction part.

How many of the users are manually labeled as spammers?

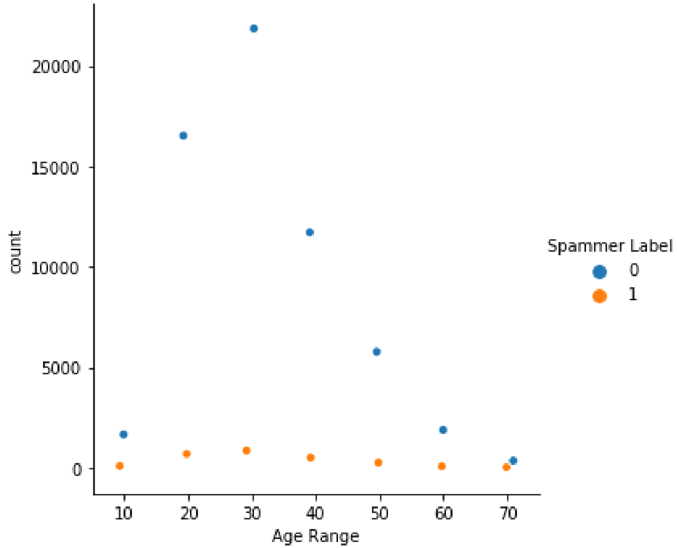
- Highly Unbalanced data
- Around 4% Spammers

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

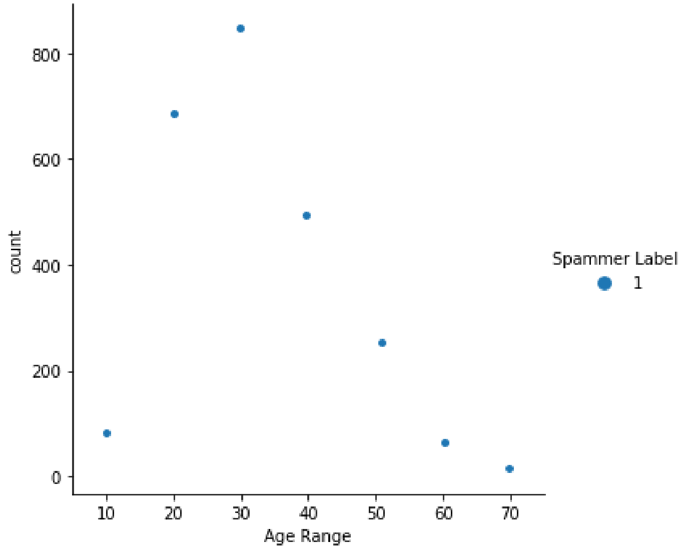
Table 1: Spammer Distribution for the Users

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

- Most of the users fall in age range of 20-50.
- Age might be a distinctive feature for Spam detection.



(a) Age Distribution for Users



(b) Spammer Age Distribution

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

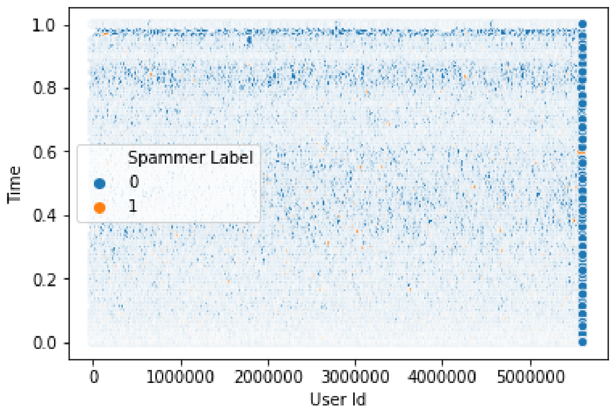
- Dominant male profile in the social network (No Gender Balance)
- Only 1.59% of the Spammers are Female.
- Gender also might be a distinctive feature for Spammer prediction.

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

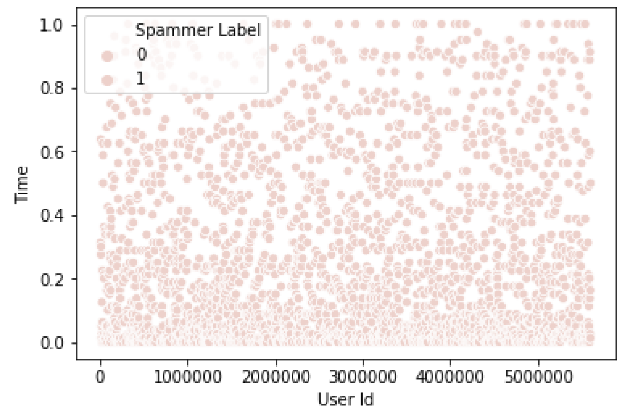
Table 2: Gender Distribution for the Users

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

- Spammers do most of their activities less than 0.2.
- Non-Spammers spent more of their time on the social network.
- Time Spent is not as distinctive as gender or age but gives us some intuition about spammers.



(a) Time Spent Plot for Users



(b) Time Spent Plot for Spammers

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

- 62055 out of 83176 relations initiated by Spammers in the data.
- Corresponds to 74.6% of the total edges.
- Remember, Spammers were only 4%.

A possible interpretation:
Majority of our dense nodes could be spammers. Outgoing and incoming edges could be an important feature as well.

What are the number of different days spammers interact?

Expectation: Spammers spam people in a short time interval with huge amount of activities.
Result: Most of the Spammers do their activities either in 1 or 2 days.

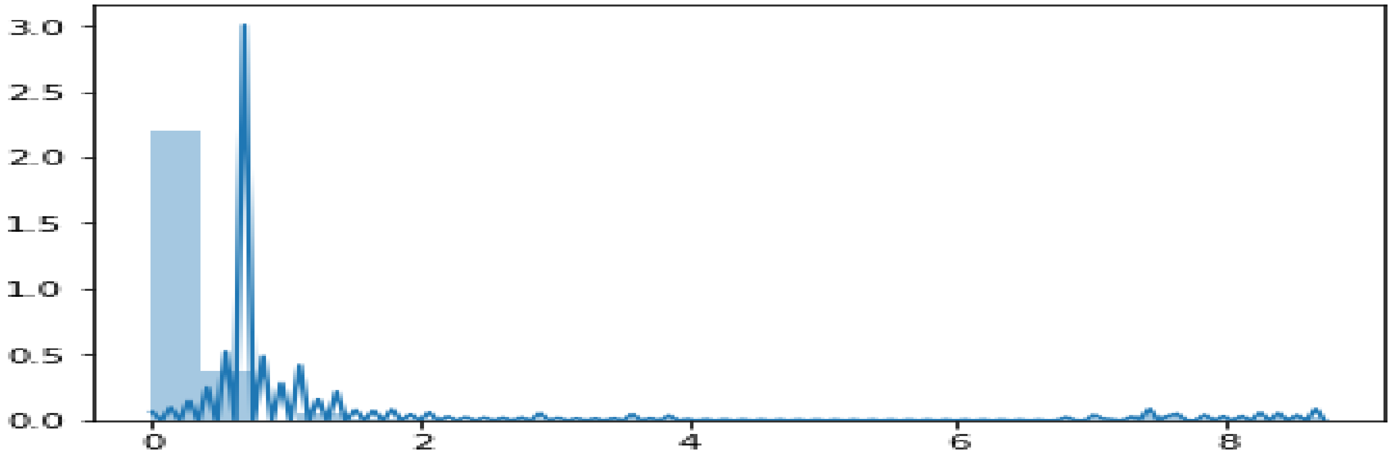


Figure 2: Different Day Numbers Distribution for Spammers

Spectral Graph Theory : Theoretical Motivations

- In Euclidean Space: Laplace Operator = Divergence of Gradient

Spectral Graph Theory : Theoretical Motivations

- In Euclidean Space: Laplace Operator = Divergence of Gradient
- What does this mean in a Graph ?

Spectral Graph Theory : Theoretical Motivations

- In Euclidean Space: Laplace Operator = Divergence of Gradient
- What does this mean in a Graph ?
    - Gradient : differences in the graph function value across each edge

Spectral Graph Theory : Theoretical Motivations

- In Euclidean Space: Laplace Operator = Divergence of Gradient
- What does this mean in a Graph ?
  - Gradient : differences in the graph function value across each edge
  - Divergence : outward flux of gradient at a given node

Graph Spectral Theory : Theoretical Motivations

- In Euclidean Space: Laplace Operator = Divergence of Gradient
- What does this mean in a Graph ?
  - Gradient : differences in the graph function value across each edge
  - Divergence : outward flux of gradient at a given node
- Higher Flux means changes in graph function vary highly from edge to edge.

Spectral Graph Theory : Theoretical Motivations

- In Euclidean Space: Laplace Operator = Divergence of Gradient
- What does this mean in a Graph ?
  - Gradient : differences in the graph function value across each edge
  - Divergence : outward flux of gradient at a given node
- Higher Flux means changes in graph function vary highly from edge to edge.
- Laplacian matrix multiplied by graph filter f: representation of curvature of f (ie. second derivative) evaluated over G

Spectral Graph Theory : Theoretical Motivations

- In Euclidean Space: Laplace Operator = Divergence of Gradient
- What does this mean in a Graph ?
  - Gradient : differences in the graph function value across each edge
  - Divergence : outward flux of gradient at a given node
- Higher Flux means changes in graph function vary highly from edge to edge.
- Laplacian matrix multiplied by graph filter f: representation of curvature of f (ie. second derivative) evaluated over G
- f projected onto L_{row_i} is curvature of f at i^{th} node in the adjacency matrix

Spectral Graph Theory : Applying to Spammers

- Spammers introduce distortions into the network :

Spectral Graph Theory : Applying to Spammers

- Spammers introduce distortions into the network :
  - More distortion means different curvatures for the same graph filter.

Spectral Graph Theory : Applying to Spammers

- Spammers introduce distortions into the network :
  - More distortion means different curvatures for the same graph filter.
  - Can these spammers be recognized by the changes in curvature they induce ?

Spectral Graph Theory : Applying to Spammers

- Spammers introduce distortions into the network :
  - More distortion means different curvatures for the same graph filter.
  - Can these spammers be recognized by the changes in curvature they induce ?
  - Can a Neural Net recognize a spammer just from the corresponding row in the Laplacian ?

Spectral Graph Theory : Applying to Spammers

- Spammers introduce distortions into the network :
  - More distortion means different curvatures for the same graph filter.
  - Can these spammers be recognized by the changes in curvature they induce ?
  - Can a Neural Net recognize a spammer just from the corresponding row in the Laplacian ?
- Neural Nets prefer uncorrelated coordinates : how do we eliminate linear dependence ?

Spectral Graph Theory : Applying to Spammers

- Spammers introduce distortions into the network :
  - More distortion means different curvatures for the same graph filter.
  - Can these spammers be recognized by the changes in curvature they induce ?
  - Can a Neural Net recognize a spammer just from the corresponding row in the Laplacian ?
- Neural Nets prefer uncorrelated coordinates : how do we eliminate linear dependence ?
  - Apply Spectral Decomposition : this will give us our features in an orthonormal space

Spectral Graph Theory : Applying to Spammers

- Spammers introduce distortions into the network :
  - More distortion means different curvatures for the same graph filter.
  - Can these spammers be recognized by the changes in curvature they induce ?
  - Can a Neural Net recognize a spammer just from the corresponding row in the Laplacian ?
- Neural Nets prefer uncorrelated coordinates : how do we eliminate linear dependence ?
  - Apply Spectral Decomposition : this will give us our features in an orthonormal space
- L_{row_i} grows with the size of the graph: can we set a hard limit on the number of features?

Spectral Graph Theory : Applying to Spammers

- Spammers introduce distortions into the network :
  - More distortion means different curvatures for the same graph filter.
  - Can these spammers be recognized by the changes in curvature they induce ?
  - Can a Neural Net recognize a spammer just from the corresponding row in the Laplacian ?
- Neural Nets prefer uncorrelated coordinates : how do we eliminate linear dependence ?
  - Apply Spectral Decomposition : this will give us our features in an orthonormal space
- $\mathrm{L}_{\mathrm{row}_i}$ grows with the size of the graph: can we set a hard limit on the number of features?
  - Apply PCA : Project $\mathrm{L}_{\mathrm{row}_i}$ onto only the first n eigenvectors

Spectral Graph Theory : Sampling

- The complete list of relations is too big to fit into memory

Spectral Graph Theory : Sampling

- The complete list of relations is too big to fit into memory
  - Manipulating the complete adjacency matrix is unrealistic, even in sparse formats

Spectral Graph Theory : Sampling

- The complete list of relations is too big to fit into memory
  - Manipulating the complete adjacency matrix is unrealistic, even in sparse formats
  - Sampling is required

Spectral Graph Theory : Sampling

- The complete list of relations is too big to fit into memory
  - Manipulating the complete adjacency matrix is unrealistic, even in sparse formats
  - Sampling is required
- Sampling technique : choose a subset of nodes uniformly , look at relations between sampled nodes

Spectral Graph Theory : Sampling

- The complete list of relations is too big to fit into memory
  - Manipulating the complete adjacency matrix is unrealistic, even in sparse formats
  - Sampling is required
- Sampling technique : choose a subset of nodes uniformly , look at relations between sampled nodes
  - Equiprobable selection of nodes and relations

Spectral Graph Theory : Sampling

- The complete list of relations is too big to fit into memory
  - Manipulating the complete adjacency matrix is unrealistic, even in sparse formats
  - Sampling is required
- Sampling technique : choose a subset of nodes uniformly , look at relations between sampled nodes
  - Equiprobable selection of nodes and relations
  - Computationally reasonable way of sampling

Spectral Graph Theory : Experiments

- Unbiased Sampling: 40 thousand training nodes, 160 thousand nodes, 100 eigenvectors

Spectral Graph Theory : Experiments

- Unbiased Sampling: 40 thousand training nodes, 160 thousand nodes, 100 eigenvectors
    - 85% accuracy on spammers / 3 % accuracy on non-spammers

Spectral Graph Theory : Experiments

- Unbiased Sampling: 40 thousand training nodes, 160 thousand nodes, 100 eigenvectors
  - 85% accuracy on spammers / 3 % accuracy on non-spammers
- Biased Sampling: 40 thousand training nodes, 1 million total nodes, 1000 eigenvectors

Spectral Graph Theory : Experiments

- Unbiased Sampling: 40 thousand training nodes, 160 thousand nodes, 100 eigenvectors
  - 85% accuracy on spammers / 3 % accuracy on non-spammers
- Unbiased Sampling: 400 thousand training nodes, 1 million total nodes, 1000 eigenvectors
  - 75% accuracy on spammers / 3% accuracy on non-spammers

Classification Method

- Artificial Neural Network
- 5 Layers(1 input, 3 hidden, 1 output)
- Total Parameters: 15,503

Dataset Used for Classification

- Original dataset consisting of 5,607,447 nodes
- Original features
- 14 additional features, based on the local centrality(in-degree, out-degree) of each node for each relation

Classification Process

- ANN
- Optimizing Function: F1 Loss
- Metrics: F1 Score
- Time to train the model took 200 Epochs with a batch size of 1,000,000 users

Classification Results

- 56-58 % on classifying Spammers
- 96-97 % on classifying Non-Spammers

Conclusion and Future Work

Thank you for your attention!