# Towards Automated Bot Detection in Political Social Networks

Berat Biçer
Vahid Namakshenas

# Motivation & Definitions

Social media impacts the political discourse and the propagation of information.

Bots threatens this for political gain in the name of lobbyists or state actors.

Bots seek to connect to many real accounts, real accounts have a mixed incoming connections.

Real accounts are local hubs, have high degree and betweenness.

# Methodology - Graph and Node-Level Analysis

**Density:** Bot clusters have larger local density than graph average.

**Clustering Coefficient (CC):** Bots likely have strong CC than network average.

**Mean path length (MPC):** MPC is likely large if bots are numerous.

**Eigenvector centrality (EIC):** Helps discovering bot farms, a cluster consists of many connected bots influencing real accounts.

**Betweenness centrality (BC):** A bot is less likely to be a broker between real accounts, and has low betweenness.

**Page Rank (PR):** A node with high Page Rank value is a reference node and is less likely to be a bot.

# Methodology - Community Detection

Bots likely share similarities and connect to other bots. So, they are likely to be placed in the same clusters.

Hierarchical clustering starting from the leaves to obtain several small communities to see whether bots can be clustered.

# Methodology - Neighbourhood Embedding

Let S be a sample in the dataset and $N_S$ be set of vertices that are the immediate neighbours of the node S.

Assume $y_S$ = 1 if S is a bot and $y_S$ = 0 otherwise.

Construct the binary string $NE_S$ where $|NE_S|$ = $||NS||$ as follows:

$$NE_S[i] = \begin{cases} 1 & if\ y_{N_S,i} = 1 \\ 0 & if\ \text{otherwise} \end{cases}$$

where $y_{Ns,i}$ is the label of *i*th neighbour of S.

$NE_S$ can be used as a feature vector, or converted into a real number for thresholding.

# Methodology - Natural Language Processing (NLP)

Learn a single representation vector for an account from a collection of tweets.

Multi-stage pipeline:

- Filtering: Removal non-English characters and emojis; hashtags, account mentions, URLs; HTML character encodings, special characters, punctuation, stop words, etc. Lemmatization.
- Vectorization: TF-IDF, word2vec[1,2], doc2vec[3], transformers[4]
- Representations merged for each account and fed to a classifier.

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.
[3] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In International conference on machine learning. PMLR, 188–1196.
[4] Yinhan Liu et al. 2019. Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

# Dataset - TwiBot-20 [1,2]

~10K labelled accounts for tweet classification, ~200 Tweets from each.

After filtering out disconnected nodes, ~2.6K accounts for network-based analysis.

Accounts are political figures & their connections from contemporary U.S. politics.

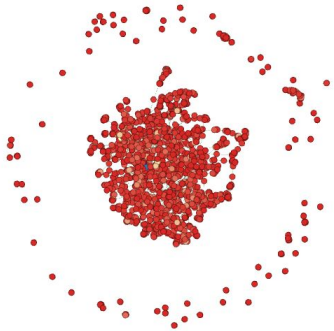


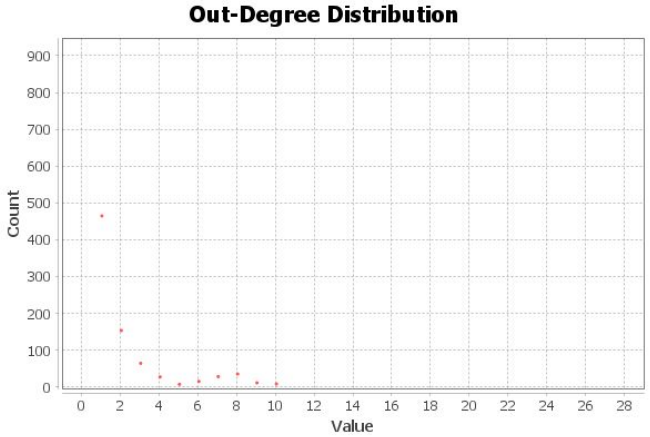
Figure 2: Layout of vertices after filtering based on degree centrality.

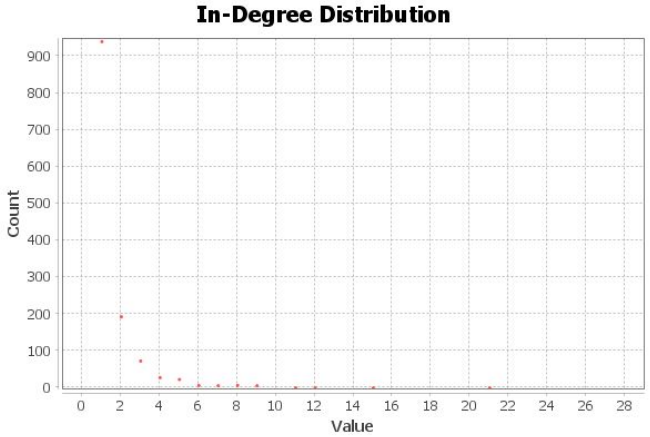
[1] Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. 2021. Twibot-20: a comprehensive twitter bot detection benchmark. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 4485–4494.
[2] Shangbin Feng et al. 2022. Twibot-22: towards graph-based twitter bot detection. arXiv preprint arXiv:2206.04564.

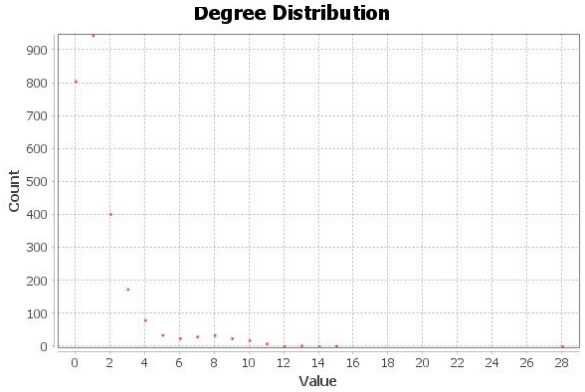
# Experiments and Results - Graph and Node-Level Analysis

| | |
|---|---|
| Nodes | 2605 |
| Links | 2070 |
| Degree Distribution | .795 |
| Density | .001 |
| Clustering Coefficient | .005 |
| Mean Path Length | 2.794 |
| Diameter | 9 |

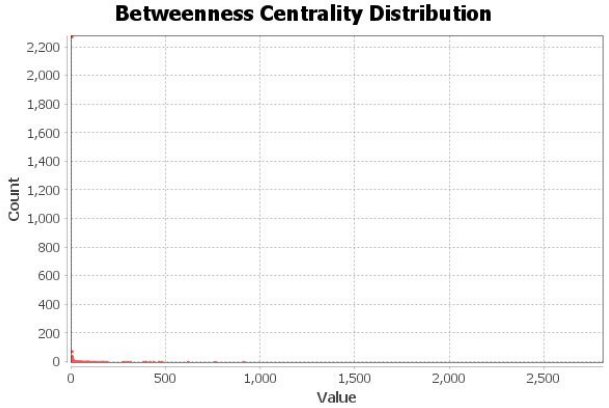
# Experiments and Results - Graph and Node-Level Analysis

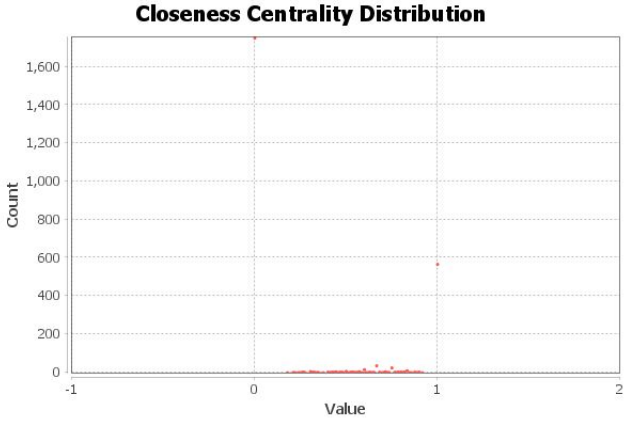


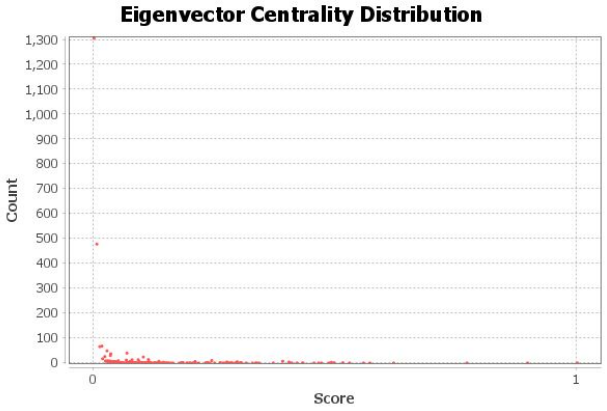



# Experiments and Results - Graph and Node-Level Analysis

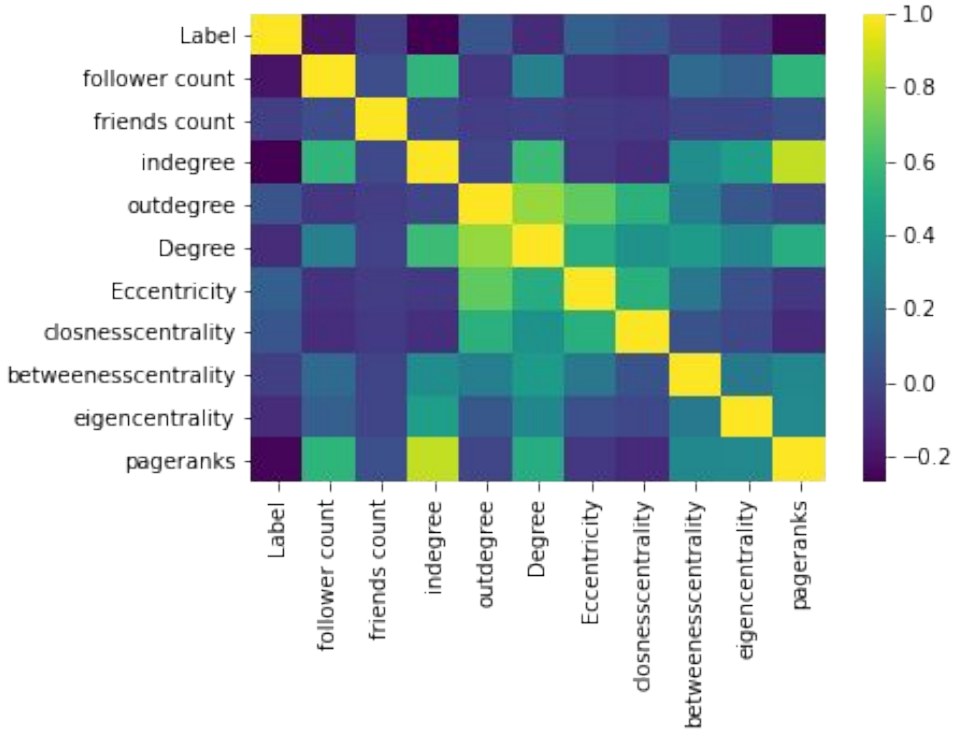


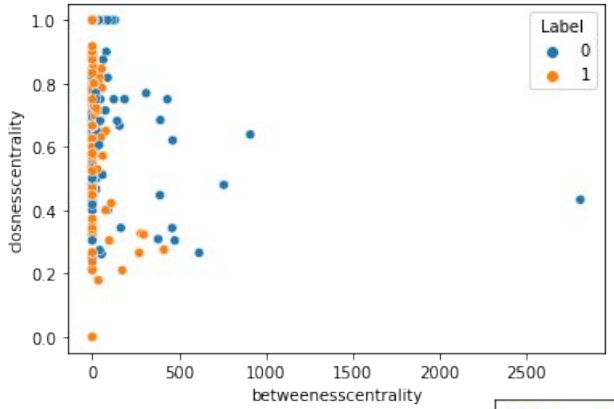



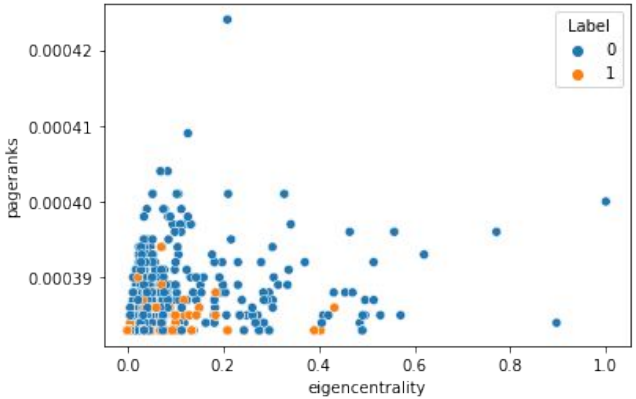
# Experiments and Results - Graph and Node-Level Analysis

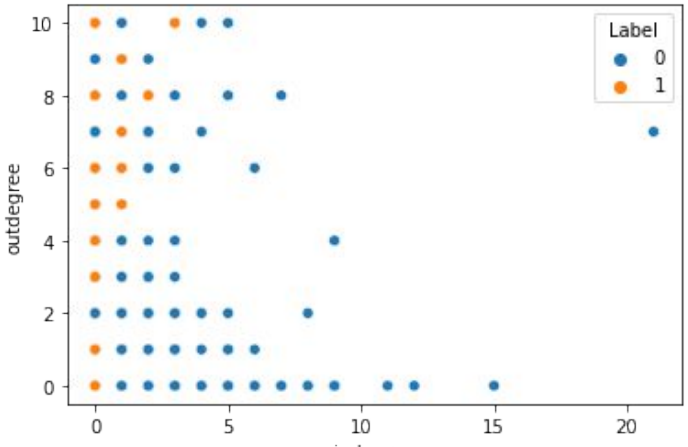
Hitmap - Correlations



# Experiments and Results - Graph and Node-Level Analysis



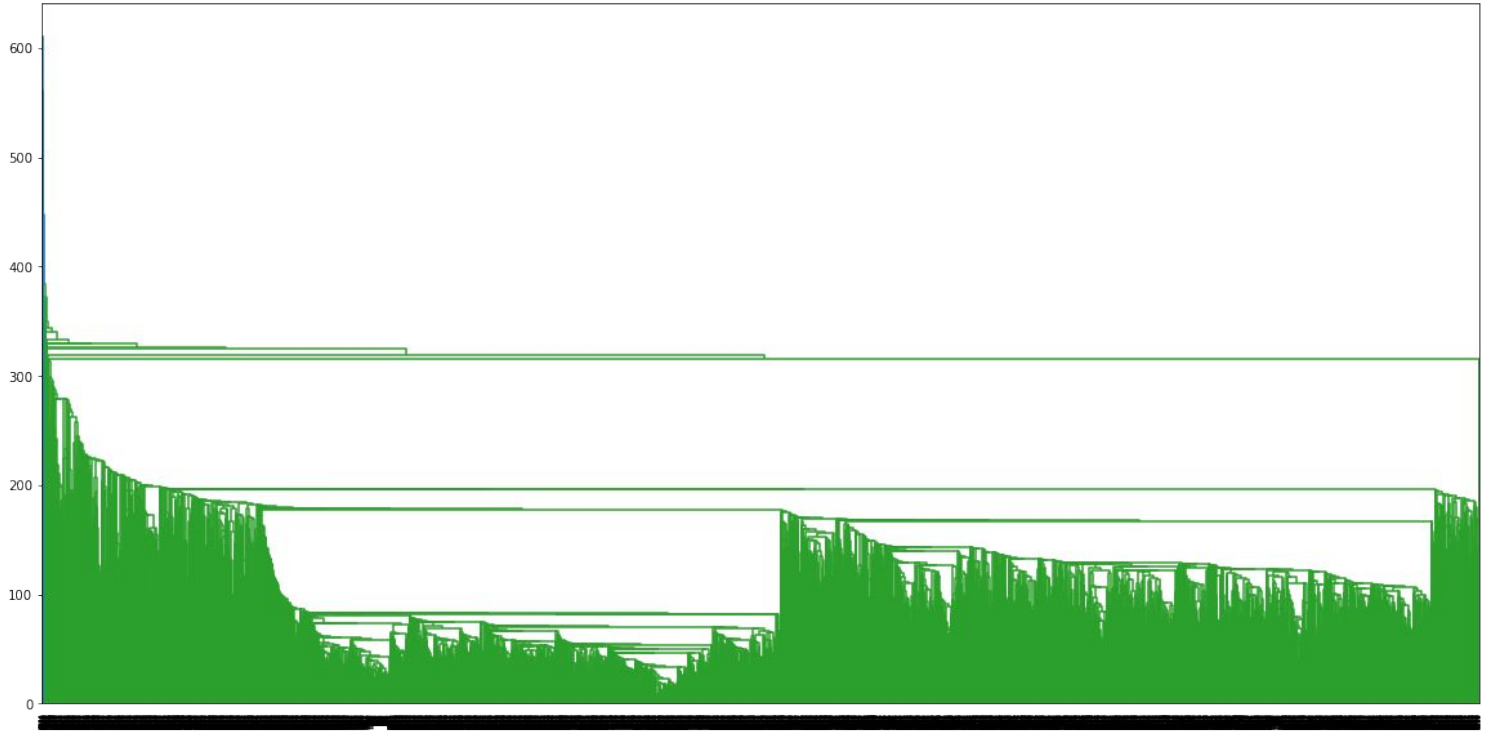
# Experiments and Results - Neighbourhood Embedding

**Table 1: Contrastive view on correct classification rate (CCR) computed over the test split. These results are obtained via experiments exploring neighbourhood information as feature representation.**

| Setup | Classifier | CCR |
|---|---|---|
| Neighbour BS | Threshold | *None* |
| Neighbour BS | Weighted Threshold | *None* |
| Neighbour Sum | Weighted Threshold | *None* |
| Neighbour Frequency | Weighted Threshold | 0.474 |
| Neighbour Embedding | MLP | *None* |
| Neighbour Embedding | SVM | *None* |
| **Neighbour Embedding** | **Random Forest** | **0.583** |

# Experiments and Results - Community Detection

Dendrogram - Hierarchical clustering



# Experiments and Results - Community Detection

Clustering Results

Bot Distribution in Clusters

| Cluster_ID | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| # of labels | 986 | 339 | 24 | 6 | 75 | 0 | 0 | 4 | 3212 |
| # of nodes | 1744 | 492 | 37 | 256 | 105 | 38 | 14 | 5 | 5585 |
| % | 56% | 69% | 65% | 2.30% | 71% | 0 | 0 | 80% | 58% |

# Experiments and Results - NLP

**Table 2: Contrastive view on correct classification rate (CCR) computed over the test split. Results are obtained when specified algorithm is used for account-level representations.**

| # Features | Method | CCR |
|---|---|---|
| 64 | doc2vec | 0.5822 |
| 128 | doc2vec | 0.5916 |
| 256 | doc2vec | 0.6036 |
| **512** | **doc2vec** | **0.6122** |
| 1024 | doc2vec | 0.5959 |
| 64 | word2vec-avg | 0.7158 |
| 128 | word2vec-avg | 0.6901 |
| **256** | **word2vec-avg** | **0.7166** |
| 512 | word2vec-avg | 0.7080 |

# Future Work

Investigating use of recurrent nets for tweet embeddings.

Investigating use of transformers for tweet embeddings.