Detecting and Visualizing Bot Prevalence across Subreddits

Team 26: Joseph Ferrin, Christopher Harmon, Samuel Parker, Joseph Rock and Mingjing Yang

Introduction

Social media bots are broadly categorized into four main types: spambots, social bots, sybil bots, and cyborgs. Spambots engage in more benign information spread. This includes spamming and advertising. Others can engage in more harmful activities such as misinformation spread, phishing, public opinion manipulation, or propaganda [new]. Hybrid or cyborg accounts operate under partial human control. Sybil bots are large groups of anonymous accounts on social media that are utilized for large effect by a single user or a small number of users.[4] The wide variety of bots along with their increased sophistication [18] has increased the need for mor sophisticated machine learning approaches.

Malicious bots give misinformation and can lead to fraud, scams, or attacks. Additionally, non-malicious bots can inadvertently contribute to the amplification of harmful content or actors overtime [1]. The motivation is to give Reddit users and researchers a clear way to see how bots behave across different communities. The tool will find accounts that are likely to act like bots, show how they post and interact within and between subreddits, and highlight the features that make them different from real people. This project will also show how bot activity overlaps between subreddits, so users can spot patterns or coordinated behavior.

Problem Definition

The aim is to help solve the problem by developing a visualization and analysis tool to identify, summarize, and compare malicious bot behavior across subreddits. We will produce per user risk scores, aggregate them to subreddit-level bot prevalence [1] and visualize cross-subreddit coordination [2] as a network whose edges represent overlap in higher-risk users. Ultimately, our work will improve transparency in social media ecosystems, providing an accessible and data driven framework for understanding automated behaviors and their impact on online discourse.

Literature Survey

On Reddit, current deduction cues include username patterns, posting frequency, and link-heavy content [3]. Supervised models are commonly used which have labeled datasets along with classification methods such as SVMs [4]. A notable study applied a network-based approach to distinguish bots and antagonistic users and compare their activity across subreddits [3]. Challenges include data collection constraints, class imbalance, non-representative datasets, and a lack of labeled datasets [4]. Semi-automatic labeling—simple heuristics plus human review—can efficiently tag large sets of social-media accounts while controlling noise, which helps build seed labels for model calibration and auditing [15]. Natural behavior on Reddit such as AMAs, live sports threads, breaking news, and daily wrap posts can create bursty patterns that trigger false positives [5]. Other approaches focus on individual accounts and struggle with coordinated bot groups [2]. Scores are usually uncalibrated with arbitrary thresholds and don’t transfer across subreddits or time [5, 2].

4. Proposed method

4.1 Overall Method and Innovations

We develop a graph-based anomaly detection framework that identifies suspicious Reddit users with the combination of network structure and user-level features. We build a co-commenter network with users as nodes and edges signifying that users commented on the same post. A unique angle to our project is that we explore using only unlabeled data. To address this challenge, we compare multiple unsupervised methods in order to perform anomaly detection and identify potential bots.

Traditional anomaly detection methods on social networks often operate only on raw user features, such as comment counts, account age, and sentiment. These methods ignore the graph structure of user interactions. Our method explores graph-derived structural signals which capture how users behave within the network, not just individually. The intuition is that anomalies - bots, spammers, sock-puppets, vote manipulators - do not behave like humans and therefore create unique graph structures. They may have unusually high interactivity if they are a spam bot posting as much as possible on multiple subreddits. While humans might develop communities based on their interests, bots don't have context and post to subreddits at random. Bots can also be deployed together to influence sentiment, which results in tight sub-networks from replying to each other. Capturing these graph features would allow us to better detect this anomalous behavior. To further capture the structure of the network, we train our model to embed the user-to-user graph as a feature vector that can be fed to anomaly detection.

Overall, our project pipeline consists of the following: 1) Querying historical subreddit data. 2) Engineer user features and create a user-to-user graph. 3) Calculate anomaly scores based on these features. 4) Display the graph of the most anomalous users to explore how they connect to different subreddits.

4.2 Data processing

For our data we use past reddit post data during May of 2015. This dataset includes over 54 million comments made during that month. To handle this amount of data, we use DuckDB to store and query the data. We clean the data by converting all the columns to the required formatting and by removing invalid users such as ‘[deleted]’.

4.3 Feature Engineering

We engineer user-level, temporal, lexical, and network features that would help to identify unusual activity.

User-level information includes number of comments, average score (upvotes and downvotes), and a suspicious username. Very high or low comment counts could signal anomalies, and low or variable scores could indicate spammy behavior. We mark usernames as suspicious if they are just random characters or if they contain bot in the name.

Temporal features such as the number of hours the user is active, the entropy of posting distribution, and time between posts. A human would be inactive for part of the day to sleep, while a bot could operate at all hours. High entropy means the posting activity is more uniform, which suggests automation. Bots would be able to post comments at a much faster rate.

Lexical features such as word diversity– the ratio of unique words to total words. Automated bots could have lower diversity due to copy-pasting and spamming. Text similarity, embedding the comments, and computing the mean similarity: Bots would be more likely to have more similarity due to repetitive phrasing or templated replies. URL link ratio, the percentage of comments that have links: Bots could be repetitively spamming links. Text length aggregation, the mean and standard deviation of comment length: Bot may have fewer variable comments if they are using fixed responses.

Network Features are constructed from a user-user graph. A network that connects users if they comment on the same post. If a network of bots is deployed, they would comment on similar posts and be more connected. This network can be embedded with other features to use for anomaly detection. The Jaccard-weighted graph [19] is normalized to reduce bias towards active users. Degree, the number of other users that someone interacts with. Bots may have high degrees if spamming or low degrees with targeted replies. The clustering coefficient measures how connected a user's local network is. Coordinated groups of bots could have high clustering.

The features are embedded along with the graph structure using Deep Graph Infomax (DGI). DGI is an unsupervised graph representation learning method that learns node embeddings by maximizing mutual information between each node's local representation and a global representation of the entire graph. The model first encodes the graph with a GCN then contrasts them against corrupted versions of the graph. Training the encoder to distinguish which nodes are real teaches meaningful structural and feature-based patterns. The resulting embedding can then be used for downstream tasks like anomaly detection.

4.3 Anomaly Detection

We use Isolation Forest to create anomaly scores for each user. Isolation Forest is an unsupervised algorithm that identifies outliers by isolating data points. It builds many random binary trees with random splits. Anomalous points tend to be isolated with fewer splits because they are located in sparse regions of the feature space. The anomaly score is created by averaging the path length of when points get isolated.

4.4 Network visualization

To showcase and explore the prevalence of bots across different subreddits, we present an interactive visualization that displays the anomalous users from two different subreddits. Users can select two subreddits to view. For each subreddit, it will display a network of the most anomalous users. These users are displayed as nodes which are colored based on their anomaly score. Hovering over a node will show the user’s features. If a user is present in both subreddits, their connection will be highlighted. This visualization can be used to show which subreddits contain more anomalous users and to what other subreddits they are connected to.

5. Evaluation

5.1 Isolation Forest and Deep Graph Infomax

Isolation forest applied to the original (untransformed) engineered features produced the strongest and most interpretable anomaly rankings. Known automated accounts and high-volume bot-like users were ranked at the top when raw graphs and behavioral metrics (degree, pagerank, betweenness, comment-length variance, score variance, etc.) were used. Applying log1p to the heavily skewed features degraded anomaly detection. Extreme values that defined many clear anomalies were compressed toward the bulk of the distribution, lowering anomaly scores for obvious bot accounts and producing a noisier top list.

Deep Graph Infomax embeddings captured structural irregularities but did not match the engineered features for identifying bot-like behavior. DGI-only rankings may have surfaced more structurally unusual users and fewer of the high-volume automated accounts. Combining DGI with engineered-feature scores added modest diversity to the candidate's set but did not materially outperform the Isolation Forest on original engineered features.

5.2 Feature Importance (PCA)

The PCA analysis shows how features contribute to variance across users. Our result show that the transformed components capture distinct behavioral dimensions overall activity levels.(PCA 1), temporal regularity (PCA 2), contribution impact (PCA 3), content verbosity (PCA 4), writing style and spam indicators (PCA 5), and network structural roles (PCA 6–8). Running Isolation Forest on the PCA-transformed data produced several high-confidence anomalies that were not surfaced by the feature-based model, including well-known automation accounts such as TrollaBot, Poem_for_your_sprog, TweetPoster, and sufficiency_bot, suggesting that PCA isolates latent behavioral axes that improve anomaly separability. The results from the PCA-transformed features and the original features show some overlap but also important differences: both methods flag clear automation accounts such as “TweetsInCommentsBot” and “ttumblrbots,” while the PCA-based analysis additionally highlights other bot-like accounts with distinctive naming or behavioral patterns, such as “TrollaBot” and “sufficiency_bot.”

Other key findings

- Super-Connectors: (PCA 1) Anomalies were structurally defined by extreme degree centrality, bridging disjointed subreddits.
- Temporal Rigidity: (PCA 2) Anomalies exhibited distinct machine-like regularity.

5.3 Evaluation on the effect of network-based features on anomaly detection

We compared Jaccard with isolation forest for this experiment. Jaccard had 31 out of the top 50 anomalies changed (62%), indicating that the model’s notion of bot behavior shifts considerably when group level structure is included. Unlike isolation forest which identifies how much a user comments and how connected they are to other users, adding Jaccard highlights users who participate in dense shared-thread neighborhoods. This is consistent with patterns expected in coordinated bot clusters. This suggests that Jaccard captures a complementary dimension of anomalous behavior. However, for a small set of manually identified suspicious users, the ranking did not significantly improve, so we treat Jaccard as a complementary network feature rather than a clearly superior method.

5.4 Weak-Label Evaluation

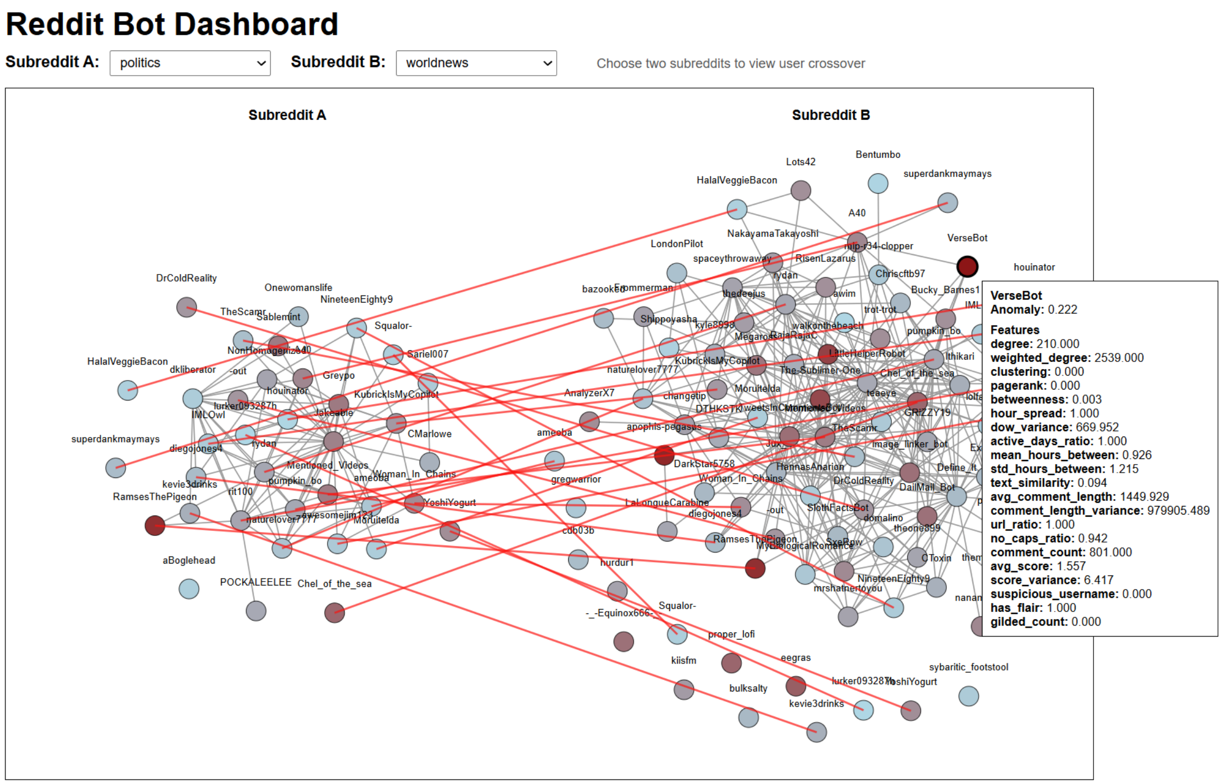
To further assess model behavior, we compared unsupervised anomaly rankings against a weak labeling heuristic based on a username pattern, user accounts containing “bot”. This heuristic identified 1,889 weak-labeled bot accounts across the dataset. Despite this, overlap with the top anomalies surfaced by the unsupervised models was very limited. The Isolation Forest trained on original engineered features shared only 13 of its top 500 anomalies with the weak-labeled set. The PCA-based Isolation Forest shared 10 of its top 500. These small overlaps indicate that most anomalous accounts detected by the model do not rely on explicit naming conventions and instead arise from behavioral or structural signals invisible to simple username rules.

Therefore, weak labels are extremely sparse and incomplete for evaluating bot detection performance.

Unsupervised models may be uncovering behavioral anomalies that weak labels fail to identify, suggesting they may have some utility in identifying non-obvious automated or coordinated accounts that would be missed by simple heuristics.

6. Conclusions and Discussion

Reddit hosts millions of users, including sophisticated bots that can evade simple detection. We developed a hybrid pipeline combining Deep Graph Infomax (DGI) and Isolation Forests to detect anomalies in a co-comment network of 632,000 users. Our unsupervised approach successfully identified “Super-Connector” bots and distinct temporal outliers, outperforming traditional keyword-based heuristics by flagging high-risk accounts that simple filters missed.



Reddit Bot Dashboard

Subreddit A: politics

Subreddit B: worldnews

Choose two subreddits to view user crossover

VerseBot
Anomaly: 0.222

Features
degree: 210.000
weighted_degree: 2539.000
clustering: 0.000
pagerank: 0.000
betweenness: 0.003
hour_spread: 1.000
dow_variance: 669.952
active_days_ratio: 1.000
mean_hours_between: 0.926
std_hours_between: 1.215
text_similarity: 0.094
avg_comment_length: 1449.929
comment_length_variance: 979905.489
url_ratio: 1.000
no_caps_ratio: 0.942
comment_count: 801.000
avg_score: 1.557
score_variance: 6.417
suspicious_username: 0.000
has_flair: 1.000
gilded_count: 0.000

Despite the strengths, some limitations remain that open opportunities for future work. The dataset is only a single month of Reddit data and it has no ground truth labels, therefore the analysis conclusion can only be drawn based on heuristics and qualitative inspections. Overall, this project demonstrates the value of integrating network analytics with unsupervised learning to spot bridge nodes (users in unrelated groups) identifying anomalies purely through their behavior.

Acknowledgement

All team members have contributed a similar amount of effort

Reference

[1] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, 2017.
[2] S. Cresci, A. Spognardi, M. Petrocchi, M. Tesconi, and R. Di Pietro, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *Companion Proceedings of the 26th International Conference on World Wide Web (WWW '17 Companion)*, 2017, pp. 963–972. doi: 10.1145/3041021.3055135.
[3] S. Hurtado, P. Ray, and R. Marculescu, "Bot detection in Reddit political discussion," in *Proceedings of the Fourth International Workshop on Social Sensing (SocialSense '19)*, New York, NY, USA: ACM, 2019, pp. 30–35. doi: 10.1145/3313294.3313386.
[4] N. Alkathiri and K. Slhoub, "Challenges in machine learning-based social bot detection: A systematic review," *Discovery Artificial Intelligence*, vol. 5, p. 214, 2025. doi: 10.1007/s44163-025-00448-w.
[5] A. Ferraz Costa, Y. Yamaguchi, A. J. M. Traina, C. Traina Jr., and C. Faloutsos, "RSC: Mining and modeling temporal activity in social media," in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*, 2015. doi: 10.1145/2783258.2783294.
[6] A. Ghosh, "Bot identification in social media," *arXiv preprint* arXiv:2503.23629, 2025. Available: https://arxiv.org/abs/2503.23629
[7] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," *arXiv preprint* arXiv:1809.10341, 2018.
[8] N. K. Ahmed and R. A. Rossi, "A web-based interactive visual graph analytics platform," *arXiv preprint* arXiv:1502.00354, 2015.
[9] D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, "Uncovering coordinated networks on social media: Methods and case studies," *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, vol. 15, no. 1, 2021. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/18075
[10] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023. doi: 10.1561/2200000101.
[11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70, 2017, pp. 1321–1330. Available: https://proceedings.mlr.press/v70/guo17a/guo17a.pdf
[12] E. Alothali, K. Hayawi, and H. Alashwal, "SEBD: A stream evolving bot detection framework with application of PAC learning approach to maintain accuracy and confidence levels," *Applied Sciences*, vol. 13, no. 7, p. 4443, 2023. doi: 10.3390/app13074443.
[13] K.-C. Yang, O. Varol, A. C. Nwala, M. Sayyadiharikandeh, E. Ferrara, A. Flammini, and F. Menczer, "Social bots: Detection and challenges," *arXiv preprint* arXiv:2312.17423, 2023. doi: 10.48550/arXiv.2312.17423.
[14] Y. Zouzou and O. Varol, "Unsupervised detection of coordinated fake-follower campaigns on social media," *EPJ Data Science*, vol. 13, art. 62, 2024. doi: 10.1140/epjds/s13688-024-00499-6.
[15] Teljstedt C, Rosell M, Johansson F. A semi-automatic approach for labeling large amounts of automated and non-automated social media user accounts. In: 2015 Second European Network Intelligence Conference; 2015. pp. 155–159.
[16] Daouadi KE, Rebaï RZ, Amous I. Bot detection on online social networks using deep forest. In: Artificial Intelligence Methods in Intelligent Algorithms: Proceedings of the 8th Computer Science On-line Conference 2019. 2019;2(8):307–315. Springer.

[17] Ding, Kaize, Jundong Li, Rohit Bhanushali, and Huan Liu. "Deep anomaly detection on attributed networks." In Proceedings of the 2019 SIAM international conference on data mining, pp. 594-602. Society for Industrial and Applied Mathematics, 2019.

[18] Aljabri, M., Zagrouba, R., Shaahid, A. et al. “Machine learning-based social media bot detection: a comprehensive literature review. Soc. Netw. Anal. Min. 13, 20 (2023). https://doi.org/10.1007/s13278-022-01020-5

[19] M. Levandowsky and D. Winter, “Distance between sets,” *Nature*, vol. 234, pp. 34–35, 1971. doi: 10.1038/234034a0.

[20] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*, 3rd ed., Cambridge University Press, Ch. 3 “Finding Similar Items.”