Towards Fair Graph Anomaly Detection: Problem, New Datasets, and Evaluation (Appendix)

A Dataset Documentation

We provide the dataset documentation by following the guidelines of the Datasheets for Datasets Gebru et al. (2021). Our code and datasets are available online.

A.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?
The datasets were created to study the `FairGAD` problem, which aims to accurately detect anomalous nodes in an input graph while avoiding biased predictions against individuals from sensitive subgroups. To the best of our knowledge, we are the first to present comprehensive and real-world ("organic") benchmark datasets that cover all of the graph, anomaly detection, and fairness aspects.

Who created the dataset (*e*.g., which team, research group) and on behalf of which entity (*e*.g., company, institution, organization)?
The datasets were created by the authors, Neng Kai Nigel Neo (M.S. Student, Georgia Tech), Yeon-Chang Lee (Post-doc, Georgia Tech), Yiqiao Jin (Ph.D. Student, Georgia Tech), Sang-Wook Kim (Professor, Hanyang University), and Srijan Kumar (Professor, Georgia Tech).

A.2 Composition

What do the instances that comprise the dataset represent (*e*.g., documents, photos, people, countries)? Are there multiple types of instances (*e*.g., movies, users, and ratings; people and interactions between them; nodes and edges)?
Each node in our `FairGAD` datasets represents a user who had posts on the Reddit or Twitter platforms. Each undirected edge in the Reddit dataset is created between two users if they post to the same subreddit within a 24-hour window, while each directed edge from user A to user B in the Twitter dataset represents user A following user B.

How many instances are there in total (of each type, if appropriate)?
There are 9,892 and 47,712 nodes, and 1,211,748 and 468,697 edges in the Reddit and Twitter datasets, respectively.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (*e*.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (*e*.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
The datasets used in our study comprise the largest connected component of each graph, consisting of all crawled users. For the Reddit dataset, we collect a list of active users in political subreddits. Here, we use a crowd-sourced collection of political subreddits[1] (see Appendix B) following previous works (Nithyanand et al., 2017). These users were collected from the period of December 10th, 2022 until March 30th, 2023. It is important to note that these users may not reflect the overall Reddit user population.

[1] https://www.reddit.com/r/redditlists/comments/josdr/list_of_political_subreddits/

Regarding the Twitter dataset, the users were obtained from a study conducted by Verma et al. (2022), which focused on analyzing COVID-19 misinformation shared among Twitter users. Consequently, the dataset is representative of users discussing misinformation related to COVID-19 on Twitter and thus may not be representative of the entire Twitter population.

By utilizing these datasets, we aim to gain insights into the `FairGAD` problem within the specific contexts of political leanings and misinformation spreaders on Reddit and Twitter, shedding light on the interplay between graph structures, anomaly detection, and fairness within these platforms.

What data does each instance consist of? "Raw" data (*e*.g., unprocessed text or images) or features? In either case, please provide a description.
Each user in the Reddit dataset is associated with data reflecting their political leanings, which was determined using the methodology outlined in Sakketou et al. (2022). Additionally, we used an average embedding of each user's all posts made in the corresponding political subreddits using the `all-MiniLM-L6-v2` model provided by Sentence-Transformers[2].

In the case of the Twitter dataset, each user possesses inferred demographic information obtained from the M3 system (Wang et al., 2019). In addition to the user's political leanings, this includes the age group (categorized as $\leq$18, 19-29, 30-39, $\geq$40), gender, and whether the Twitter account is an organization account. Additionally, the number of favorites and the account verification status were recorded. Similar to the Reddit dataset, the users' post histories were retrieved and embedded using the multilingual model `multi-qa-MiniLM-L6-cos-v1`.

Is there a label or target associated with each instance?
Yes, each user is assigned a label of 1 if the user has a higher frequency of posting misinformation links compared to real news links, and 0 otherwise. The determination of these links was based on domain annotations provided by Sakketou et al. (2022).

Is any information missing from individual instances?
No.

Are relationships between individual instances made explicit (*e*.g., users' movie ratings, social network links)?
Yes. As mentioned above, each undirected edge in the Reddit dataset is created between two users if they post to the same subreddit within a 24-hour window, while each directed edge from user A to user B in the Twitter dataset represents user A following user B.

Are there recommended data splits (*e*.g., training, development/validation, testing)?
No, we treat the `FairGAD` problem as an unsupervised learning task, where the entire graph is utilized for training without any labeled data. In addition, it is worth noting that existing GAD methods identify anomalies based on the top nodes with high reconstruction errors, without the need for a separate classifier to be trained. Therefore, there is no need for specific data splits in this context.

Are there any errors, sources of noise, or redundancies in the dataset?
Yes, partially. As user demographic information is inferred for users in the Twitter dataset, we expect some noise in this category of data.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (*e*.g., websites, tweets, other datasets)?
Yes, our `FairGAD` datasets are self-contained.

Does the dataset contain data that might be considered confidential (*e*.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public

[2] `https://www.sbert.net`

communications)?
All of the raw data in the datasets are derived from publicly available sources. For Reddit, the data are directly accessible through the Pushshift API or Reddit dumps[3] and are also accessible through the PRAW API[4]. In the case of Twitter, the data can are directly accessible using registered Twitter API[5].

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
No, all of the user postings are represented as low-dimensional embeddings.

Does the dataset identify any subpopulations (*e*.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
Yes. The assignment of political leanings (*i*.e., sensitive attribute) is determined using the methodology outlined in Sakketou et al. (2022). Users are classified as right-leaning (with a sensitive label of 1) if they post a higher number of links from right-leaning sites than left-leaning sites, and as left-leaning (with a sensitive label of 0) in the opposite case. As a result, approximately 13% and 12% of the users were identified as having a positive sensitive label, indicating a right-leaning political affiliation, for the Reddit and Twitter datasets, respectively.

The assignment of misinformation spreaders (*i*.e., anomaly label) is also determined similarly to the political leanings. Users are classified as misinformation spreaders (with an anomaly label of 1) if they have a higher frequency of posting misinformation links compared to real news links, and as real news spreaders (with an anomaly label of 0) in the opposite case. As a result, approximately 14% and 7% of the users were identified as having a positive anomaly label, indicating the misinformation spreaders, for the Reddit and Twitter datasets, respectively.

Regarding other subpopulations such as age and gender, the user demographic information for the Twitter dataset was inferred using the M3 System (Wang et al., 2019). As a result, the dataset consists of 47.7% users $\geq$40 years old, 20.6% between 30-39 years old, 22.2% between 19-29 years old, and 9.6% $\leq$18 years old. As for gender, it includes that 73% of users are male and 27% of users are female.

Is it possible to identify individuals (*i*.e., one or more natural persons), either directly or indirectly (*i*.e., in combination with other data) from the dataset?
No, it is not possible to identify individuals from the datasets as usernames are not included in the datasets.

Does the dataset contain data that might be considered sensitive in any way (*e*.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?
Yes, our `FairGAD` datasets contain data that might be considered sensitive. Based on the methodology outlined in Sakketou et al. (2022), the political affiliation and anomaly label of users were inferred from publicly posted information.

A.3 Collection Process

What mechanisms or procedures were used to collect the data (*e*.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

[3] `https://pushshift.io/`

[4] `https://praw.readthedocs.io/en/stable/index.html`

[5] `https://developer.twitter.com/en/products/twitter-api`

We used the Pushshift API[6] to collect Reddit data, and the Twitter API[7] to collect Twitter data. All the scripts we used for data collections were written in Python.

Who was involved in the data collection process (*e*.g., students, crowdworkers, contractors) and how were they compensated (*e*.g., how much were crowdworkers paid)?
All data collection process was done by the authors.

Over what timeframe was the data collected?
The Reddit dataset encompasses the entire history of Reddit from its inception in June 2005 until March 2023. For the Twitter dataset, the user data was retrieved from Verma et al. (2022) and collected over a period of time from January 1, 2019 to July 15, 2020. The entire data collection process started on December 10, 2022 and ended on March 30, 2023.

Were any ethical review processes conducted (*e*.g., by an institutional review board)?
The collection of publicly available datasets was determined to be review exempt by the Institutional Review Board (IRB). Furthermore, we will strictly adhere to the privacy policies of Twitter[8] and Reddit[9] to ensure the protection of personally identifiable information when releasing our datasets.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (*e*.g., websites)?
The datasets were collected directly from the Reddit and Twitter platforms.

Were the individuals in question notified about the data collection?
No.

Did the individuals in question consent to the collection and use of their data?
No, as all the information collected is publicly available from the users' posts. The data collection process followed the privacy policies of Reddit[10] and Twitter[11].

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?
N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (*e*.g., a data protection impact analysis) been conducted?
N/A.

A.4 Preprocessing/Cleaning/Labeling

Was any preprocessing/cleaning/labeling of the data done (*e*.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?
Political affiliation was done by counting the number of left-leaning and right-leaning news domain links for each user, and misinformation similarly done by counting the number of misinformation and real news domain links, with news domains annotated by Sakketou et al. (2022). Embedding was done for each user's posts using Sentence Transformers Reimers & Gurevych (2019). User demographic information for the Twitter dataset was done using the M3 system Wang et al. (2019). We did not perform dataset cleaning such as removal of instances.

[6] https://pushshift.io/

[7] https://developer.twitter.com/en/products/twitter-api

[8] https://twitter.com/en/privacy

[9] https://www.reddit.com/policies/privacy-policy

[10] https://www.reddit.com/policies/privacy-policy

[11] https://twitter.com/en/privacy

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (*e*.g., to support unanticipated future uses)?
Yes.

Is the software that was used to preprocess/clean/label the data available?
We plan to release the code and datasets.

A.5 USES

Has the dataset been used for any tasks already?
No, our `FairGAD` datasets were collected and curated to study the `FairGAD` problem for the first time. In this work, we evaluated the effectiveness and limitations of existing GAD methods with fairness methods in addressing the `FairGAD` problem.

What (other) tasks could the dataset be used for?
The datasets could be used as additional datasets for research on fair graph mining or graph anomaly detection.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (*e*.g., stereotyping, quality of service issues) or other undesirable harms (*e*.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
The Reddit dataset was collected using the Pushshift API. Notably, Pushshift will no longer ingest new content from Reddit starting in May 2023. If prospective users wish to gather more attributes associated with our datasets, they may need to explore alternative packages like PRAW[12]. For the Twitter dataset, the information was collected using Twitter's API as of March 2023, which will require a change in the future as the API is updated for further collection.

Meanwhile, we will guarantee the availability of our datasets to the research community.

Are there tasks for which the dataset should not be used?
The datasets should not be used to identify political affiliation or misinformation spreaders in domains unrelated to politics or COVID-19, as the sample of users in these datasets is not representative of all users on the respective social media platforms.

A.6 DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (*e*.g., company, institution, organization) on behalf of which the dataset was created?
Yes, the dataset will be freely available for distribution once accepted.

How and When will the dataset be distributed (*e*.g., tarball on website, API, GitHub)?
The dataset will be distributed with a publicly accessible link.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
The dataset will be licensed under the BSD-3 Clause license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
To the best of our knowledge, no third parties have imposed IP-based or other restrictions on the dataset.

[12] `https://praw.readthedocs.io/en/stable/index.html`

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?
To the best of our knowledge, there are no export controls or regulatory restrictions applicable to the dataset or individual instances.

A.7 Maintenance

How can the owner/curator/manager of the dataset be contacted (*e*.g., email address)?
The authors can be contacted via their email or via GitHub issues.

Is there an erratum?
Erratas will be posted on the GitHub repository, along with version changes.

Will the dataset be updated (*e*.g., to correct labeling errors, add new instances, delete instances')? If so, please describe how often, by whom, and how updates will be communicated to users (*e*.g., mailing list, GitHub)?
Yes, as errors or relevant information is identified by the authors, new versions of the datasets will be made publicly available. Updates will be communicated through the GitHub repository.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (*e*.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?
No specific limits on data retention associated with the instances are applicable.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
Yes, all versions of the datasets will be hosted and supported.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
Errors and notes can be submitted via GitHub issues on the GitHub repository, where the authors will verify and discuss the submissions before updating the datasets.

B List of Politics Related Subreddits

We used a crowd-sourced collection of political subreddits[13] following previous works (Nithyanand et al., 2017).

“r/politics", “r/Liberal", “r/Conservative", “r/Anarchism", “r/LateStageCapitalism", “r/PoliticalDiscussion", “r/PoliticalHumor", “r/worldpolitics", “r/PoliticalCompassMemes", “r/PoliticalVideo", “r/PoliticalDiscourse", “r/PoliticalFactChecking", “r/PoliticalRevisionism", “r/PoliticalIdeology", “r/PoliticalRevolution", “r/PoliticalMemes", “r/PoliticalModeration", “r/PoliticalCorrectness", “r/PoliticalCorrectnessGoneMad", “r/PoliticalTheory", “r/PoliticalQuestions", “r/PoliticalScience", “r/PoliticalHumorModerated", “r/PoliticalCompass", “r/PoliticalDiscussionModerated", “r/worldnews", “r/news", “r/worldpolitics", “r/worldevents", “r/business", “r/economics", “r/environment", “r/energy", “r/law", “r/education", “r/history", “r/PoliticsPDFs", “r/WikiLeaks", “r/SOPA", “r/NewsPorn", “r/worldnews2", “r/AnarchistNews", “r/republicofpolitics", “r/LGBTnews", “r/politics2", “r/economic2", “r/environment2", “r/uspolitics", “r/AmericanPolitics", “r/AmericanGovernment", “r/ukpolitics", “r/canada", “r/euro", “r/Palestine",

[13] https://www.reddit.com/r/redditlists/comments/josdr/list_of_political_subreddits/

“r/eupolitics", “r/MiddleEastNews", “r/Israel", “r/india", “r/pakistan", “r/china", “r/taiwan", “r/iran", “r/russia", “r/Libertarian", “r/Anarchism", “r/socialism", “r/progressive", “r/Conservative", “r/americanpirateparty", “r/democrats", “r/Liberal", “r/new_right", “r/Republican", “r/egalitarian", “r/demsocialist", “r/LibertarianLeft", “r/Liberty", “r/Anarcho_Capitalism", “r/alltheleft", “r/neoprogs", “r/democracy", “r/peoplesparty", “r/Capitalism", “r/Anarchist", “r/feminisms", “r/republicans", “r/Egalitarianism", “r/anarchafeminism", “r/Communist", “r/socialdemocracy", “r/conservatives", “r/Freethought", “r/StateOfTheUnion", “r/equality", “r/propagandaposters", “r/SocialScience", “r/racism", “r/corruption", “r/propaganda", “r/lgbt", “r/feminism", “r/censorship", “r/obama", “r/war", “r/antiwar", “r/climateskeptics", “r/conspiracyhub", “r/infograffiti", “r/CalPolitics", “r/politics_new"

REFERENCES

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, 2021.

Rishab Nithyanand, Brian Schaffner, and Phillipa Gill. Online political discourse in the trump era. *CoRR*, abs/1711.05303, 2017. URL `http://arxiv.org/abs/1711.05303`.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. In *EMNLP*, 2019.

Flora Sakketou, Joan Plepi, Riccardo Cervero, Henri Jacques Geiss, Paolo Rosso, and Lucie Flek. FACTOID: A new dataset for identifying misinformation spreaders and political bias. In *LREC*, pp. 3231–3241, 2022.

Gaurav Verma, Ankur Bhardwaj, Talayeh Aledavood, Munmun De Choudhury, and Srijan Kumar. Examining the impact of sharing covid-19 misinformation online on mental health. *Scientific Reports*, 12(1):1–9, 2022.

Zijian Wang, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flöck, and David Jurgens. Demographic inference and representative population estimates from multilingual social media data. In *WWW*, pp. 2056–2067, 2019.