Fairness Through Controlled (Un)Awareness in Node Embeddings

Dennis Vetter [*1] **Jasper Forth** [*2] **Gemma Roig** [1] **Holger Dell** [2]

Abstract

Graph representation learning is central for the application of machine learning (ML) models to complex graphs, such as social networks. Ensuring 'fair' representations is essential, due to the societal implications and the use of sensitive personal data. In this paper, we demonstrate how the parametrization of the *CrossWalk* algorithm influences the ability to infer a sensitive attributes from node embeddings. By fine-tuning hyperparameters, we show that it is possible to either significantly enhance or obscure the detectability of these attributes. This functionality offers a valuable tool for improving the fairness of ML systems utilizing graph embeddings, making them adaptable to different fairness paradigms.

1. Introduction

The expanding application of machine learning (ML) systems across various sectors often leads to disparate impact on different subgroups, frequently disadvantaging minorities or marginalized groups (Barocas & Selbst, 2016; Angwin et al., 2016; Angwin & Parris, 2016; Kaplan et al., 2022). In the context of social networks, these ML systems handle highly personal data and risk negatively impacting lives (Angwin & Parris, 2016; Mello-Klein, 2022; Kaplan et al., 2022; Hao, 2021).

A common pre-processing step in ML pipelines for large and complex graphs, like social networks, is representation learning, where each node is mapped to a fixed-dimensional real vector that captures relevant structural properties (Xu, 2021). Methods for learning node representations include various approaches, with random walks being a common technique exemplified by *DeepWalk* (Perozzi et al., 2014) and *node2vec* (Grover & Leskovec, 2016). *DeepWalk* utilizes uniformly random walks to capture network structures while *node2vec* introduces hyperparameters to adapt the walk strategies for better embeddings in different configurations (Xu, 2021).

Khajehnejad et al. (2022) introduced *CrossWalk* to make the node embeddings generated by *node2vec* fairer. In particular, the input vertices are assumed to have *sensitive attributes*, such as *gender* or *ethnicity*, and the random walk is biased to walk towards group boundaries. This introduces a trade-off between the fairness properties of the node embeddings and the ability for the node embeddings to accurately capture the original graph structure. However, the authors do not provide clear guidelines for parameterizing their algorithm and do not evaluate fairness beyond the disparity metric. In our work, we address these limitations and evaluate *CrossWalk* using various parametrizations and under different notions of fairness. Our contributions are threefold:

- We demonstrate how *CrossWalk* can be parameterized to adjust the detectability of sensitive attributes from node embeddings, making it either more difficult or easier to recover these attributes.
- We analyze the impact of fairness improvements on embedding quality by including a non-sensitive control attribute in our investigations.
- We provide an integrated implementation to apply and evaluate both *CrossWalk* and *node2vec*, enabling easy comparison of their embeddings.

2. Related Work

2.1. Fairness

In the machine learning literature, the terms 'fair' or 'just' are often used interchangeably, but they have different meanings in moral philosophy: *Fairness* is subjective and dependent on personal perception, whereas *justice* is objective, relying on external standards such as laws or other rules. This distinction is relevant when defining ML 'fairness metrics', which do not measure subjective fairness, but align more with the philosophical concept of justice (Kordzadeh & Ghasemaghaei, 2022; Colquitt & Rodell, 2015; Goldman & Cropanzano, 2015).

Fairness in ML typically involves grouping the population based on 'sensitive attributes' like ethnicity, gender, or age

[*]Equal contribution [1]Computational Vision and Artificial Intelligence Lab, Goethe University Frankfurt, Frankfurt am Main, Germany [2]Theoretical Computer Science Group, Goethe University Frankfurt, Frankfurt am Main, Germany. Correspondence to: Dennis Vetter <vetter@em.uni-frankfurt.de>.

ICML 2024 workshop on the Next Generation of AI Safety.
Copyright 2024 by the author(s).

(Mehrabi et al., 2022; Barocas & Selbst, 2016; Kuppler et al., 2021). In our work, we focus on two concepts for fair ML: *procedural justice* and *distributive justice*. Procedural justice concerns just processes (Rawls, 1999; Miller, 2023), aiming to prevent ML systems from using sensitive attributes (Kusner et al., 2017; Grgic-Hlača et al., 2016). This approach faces challenges: (1) the system should not infer group membership from non-sensitive attributes, and (2) excluding data might worsen results (Barocas & Selbst, 2016; Grgic-Hlača et al., 2016; Kozyrkow, 2023).

In contrast, distributive justice focuses on the just allocation of outcomes (Rawls, 1999; Miller, 2023), ensuring similar performance across sub-groups (Mehrabi et al., 2022; Kordzadeh & Ghasemaghaei, 2022; Kuppler et al., 2021). The choice between procedural and distributive justice depends on the use-case, highlighting the need for a holistic approach in ethical applications of ML (Kordzadeh & Ghasemaghaei, 2022; Colquitt & Rodell, 2015; AI HLEG, 2019).

2.2. Learning Graph Representations

In the context of our work, the goal of representation learning is to create an *embedding*, mapping graph nodes to fixed-dimensional real vectors while preserving node relations and features (Perozzi et al., 2014; Grover & Leskovec, 2016). This enables the application of standard statistical and ML approaches. Node similarity can be defined by *homophily* – similar neighbors, or *structural equivalence* – similar neighborhood structures.

A common method for learning embeddings uses the *Skip-Gram* model (Mikolov et al., 2013), often applied in natural language processing. In this analogy, nodes are 'words' and walks are 'sentences' (Perozzi et al., 2014; Tang et al., 2015; Grover & Leskovec, 2016). Under this model, the strategy that produces the walks is important.

DeepWalk performs random walks by uniformly sampling nodes and their neighbors (Perozzi et al., 2014). *node2vec* generalizes the sampling process by using edge weights as transition probabilities and introducing two hyperparameters p and q. The parameter p controls exploration away from the walk's root, while q balances between homophily and structural equivalence (Grover & Leskovec, 2016).

CrossWalk (Khajehnejad et al., 2022) employs a fairness intervention before learning representations with *node2vec*. It groups nodes by a sensitive attribute and adjusts transition probabilities to favor crossing group boundaries, reducing the distance in embedding space between nodes from different groups. CrossWalk uses two hyperparameters, α and β.[1] The parameter α controls the likelihood of crossing group boundaries, while β encourages visits near group boundaries.

[1]Khajehnejad et al. (2022) use the names α and p, we changed p to β to avoid confusion with *node2vec*'s p.

3. Experimental setup

3.1. Datasets

Obtaining diverse graph datasets with multiple attributes per node proved challenging. To adress this, we utilized demographic-rich data from the Pokec social network and created multiple controlled subgraph datasets to evaluate *CrossWalk* in different settings.

The base of our datasets is the Pokec social network dataset, a collection of anonymized user data, including attributes like age, place of living, and connections with other users (Takac & Zabovsky, 2012), sourced from SNAP (Leskovec & Krevl, 2014). The total graph comprises over 1.6 million nodes and over 30 million edges (Takac & Zabovsky, 2012).

In our experiments, we focus on the attributes *age* and *location*, and we select only nodes where these attributes are available. Preliminary analysis revealed that nodes with similar attributes tend to be more densely connected.

We define three categories of subgraphs: (1) *distinct*: small towns characterized by few connections between each other, due to their geographical separation; (2) *semi-distinct*: clusters of well-connected small towns in geographic proximity, with few connections between geographically distant groups; and (3) *mixed*: adjacent city districts within the same urban area.

Table 1. Summary of the experiment data. Each category consists of three subgraphs with varying numbers of nodes, edges, and location groups.

	Sub-Graph	#Nodes	#Edges	#Groups Location	#Groups Age
Distinct	0	2,504	10,685	3	3
	1	20,927	127,329	4	3
	2	24,228	145,488	8	3
Semi	0	9,240	33,961	4	3
	1	32,640	183,478	12	3
	2	35,086	232,304	14	3
Mixed	0	15,677	26,339	5	3
	1	32,973	180,788	8	3
	2	49,222	138,405	12	3

For each category, we compile three different datasets with manually selected locations based on geographical proximity. This approach allowed us to control the number of connections between nodes from different locations, creating 'easier' (distinct category) and 'harder' (mixed category) node classification problems. We observed that locations are internally well-connected, and geographically close locations share more connections than geographically distant ones. An example of this can be seen in Figure 6 in the appendix, which illustrates a graph from the *semi-*

distinc category. This pattern of internal connectivity and proximity-based connections holds across all other datasets.

In addition, we categorize the age attribute into three groups: (a) 16-18 years, (b) 19-21 years, and (c) 22 years and older, to obtain differing group sizes and approximately equal levels of interconnectedness that are more stable across the different subgraph categories. With this, we can analyze how the performance of the *CrossWalk* algorithm changes under different pre-conditions, and how the generated representations perform with respect to a non-sensitive attribute. Table 1 summarizes the different controlled subgraph selections used in our experiments.

3.2. Experimental Pipeline

Our experimental pipeline is structured as follows:

1. *Dataset selection:* We choose a subgraph from one of the three categories: *distinct*, *semi-distinct*, or *mixed*.
2. *CrossWalk biasing:* We select parameters (α and β) and a sensitive attribute ('location' or 'age') to bias the transition probabilities within the graph.
3. *node2vec embedding:* We run the *node2vec* algorithm with a set of parameters (p and q) to generate node embeddings through a collection of random walks.
4. *Attribute prediction:* From the resulting node embeddings, we predict either the sensitive attribute or a control attribute ('location' or 'age') using the Label Propagation algorithm (Zhu & Ghahramani, 2002) with 50% of nodes labeled. We employed 25-fold cross-validation for robust performance estimation.

The effect of the hyperparameters of *node2vec* (p, q) and *CrossWalk* (α, β) is summarized in Table 2. We evaluated all combinations of the hyperparameters with the following values:

- p, q: 0.1, 0.5, 1, 5.0, 10.0
- α: 0.01, 0.25, 0.5, 0.75, 0.99
- β: 1, 2, 3, 5, 8, 11, 15

We further alternate between using *location* and *age* as the sensitive attributes to assess the adaptability of *Crosswalks*'s fairness intervention. The source code to create controlled subgraph datasets, to bias the transition probabilities, to generate the random walks, and to perform the experiments is available under an open-source license.[2]

[2] https://github.com/jasperforth/fairgraphwalk-experiments

Table 2. Effect of the hyperparameters in *node2vec* (p, q) and *CrossWalk* (α, β).

Param.	Effect
p	Controls walk distance from root. $p < \min(q, 1)$ for local revisits $p > \max(q, 1)$ for exploration
q	Controls notion of similarity. $q < 1$ for DFS style (homophily) $q > 1$ for BFS style (structural equivalence)
α	Controls group boundary crossing. Higher values make crossing group boundaries more likely.
β	Encourages walks towards nodes on group boundaries. Higher values correspond to stronger encouragement.

3.3. Evaluation Method

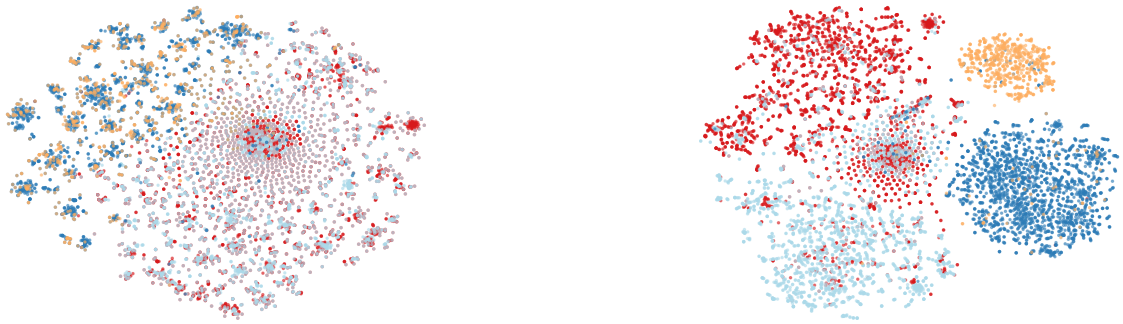
Our evaluation focuses on three key aspects: (i) awareness, (ii) disparity, and (iii) classification performance. The baseline for our analysis consists of embeddings created by *node2vec* without *CrossWalk* biasing. Since *node2vec* embeddings are produced without using node attributes, this baseline helps us analyze the effect of *CrossWalk*'s fairness intervention.

Let $G = (V, E)$ be a graph where each node $v \in V$ is associated with a sensitive attribute and a control attribute, and the nodes are partitioned into C distinct groups $1, \ldots, C$ based on the sensitive attribute. For each group $i \in \{1, \ldots, C\}$ defined by the sensitive attribute, we write $Q_i \in [0, 1]$ for the ability to predict it, measured as F1 score of Label Propagation, and let $Q_i^* \in [0, 1]$ be the corresponding prediction F1 score for the control attribute, based on the same sensitive attribute groups. We use the following metrics in our evaluation:

Awareness. Defined as $\max(Q_i \colon i \in \{1, \ldots, C\})$, this metric captures the ability to infer group membership from the embeddings. Lower *awareness* indicates better procedural justice.

Disparity. Following Khajehnejad et al. (2022), we measure *disparity* as $\mathrm{var}(Q_i \colon i \in \{1, \ldots, C\})$. This metric compares the classification F1 score across different groups, with lower disparity indicating better distributive justice.

Performance. We use *performance* as $\mathrm{mean}(Q_i^* \colon i \in \{1, \ldots, C\})$, to measure the ability of correctly predicting the control attribute across sensitive attribute groups. If *performance* stays high regardless of whether the *CrossWalk* intervention was used before applying *node2vec*, then applying *CrossWalk* does not degrade the quality of embeddings for non-sensitive attributes.



(a) The *low awareness* configuration improves procedural justice, as inferring the sensitive attribute is more difficult.

(b) The *high awareness* configuration improves distributive justice, as inferring the sensitive attribute is easier for all groups.

Figure 1. Impact of *low awareness* and *high awareness* configurations of the *CrossWalk* algorithm on node embeddings for a subgraph from the *semi-distinct* category, with sensitive attribute 'location'. The embeddings are visualized using t-SNE for dimensionality reduction, and the node colors correspond to sensitive attribute class.

4. Results

This section presents the impact of *node2vec* and *CrossWalk* parametrizations on our evaluation metrics: awareness, disparity, and classification performance.

Our experiments show that the *node2vec* parametrization alone does not significantly influence our evaluation metrics. This can be seen in Figure 2, where the error bars indicate the value range across all evaluated combinations of *node2vec* hyperparameters. In contrast, the *CrossWalk* parameters have a much higher impact on the representations. We identify two main configurations: *low awareness* and *high awareness*, illustrated by Figure 1.

Low awareness is achieved through high values for the hyperparameters α and β. In our experiments this corresponds to $\alpha = 0.99$ and $\beta = 15$. With this configuration, the random walks are biased towards visiting nodes at group boundaries and frequently crossing group boundaries. This leads to embeddings where the groups are much more different to separate. This intermixing of embeddings is more effective when the groups are well interconnected.

Figure 1 shows the resulting embeddings for the same dataset from the *semi-distinct* category under the two different configurations. Figure 2 indicates that these results are consistent across the three graph categories. In the *low awareness* configuration depicted in Figure 1(a), embeddings from geographically close locations, such as yellow and dark blue, are much more intermixed than the embeddings of nodes from geographically distant locations, such as yellow and red. Another important observation is that while *CrossWalk* is able to mix the embeddings from the geographically close locations, in none of our evaluated configurations were the embeddings from geographically distant locations intermixed. A possible cause could be that there are simply not enough interconnections between the groups that allow biased random walks to cross these specific group borders.

High awareness is achieved through low values for the hyperparameters of α and β. In our experiments, this corresponds to $\alpha = 0.01$ and $\beta = 1$. This configuration discourages the crossing of boundaries and biases the random walks towards staying in the same group as the start node. This results in similar embeddings for nodes from the same group and different embeddings for nodes from different groups, as seen in Figure 1(b).

In addition, it can be seen that the *CrossWalk* intervention makes it possible to increase awareness above or below the *node2vec* baseline.

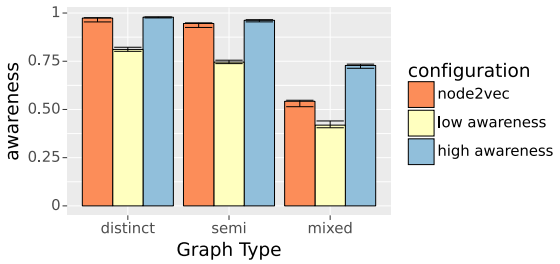


Figure 2. Mean awareness over subgraphs for sensitive attribute 'location'. Error bars show range of values over all *node2vec* parametrisations. *Low* and *high awareness* configurations can consistently adapt awareness below or above the *node2vec* baseline.

We also found that the *low awareness* configuration tends to increase the *disparity* metric, which is highlighted in Figure 4. The combination of reduced awareness and increased disparity between the *low awareness* and *high awareness* configurations suggests that the changes in classification score are not equally distributed. As seen in Figure 3, this

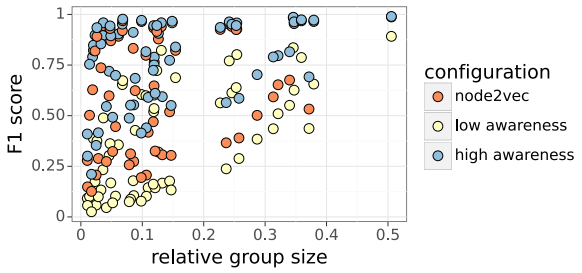


Figure 3. Classification F1 score for sensitive attribute 'location' by relative group size of groups in all subgraphs. *Low awareness* makes classification of small groups virtually impossible, differences in F1 score between *low* and *high awareness* are more pronounced for smaller groups.

difference tends to be more pronounced for smaller groups.

In the *high awareness* configuration, the classification F1 score across groups becomes more equal, leading to reduced *disparity* (see Figure 3) and improved distributive justice. Conversely, in the *low awareness* configuration, it can become virtually impossible to correctly classify members of underrepresented groups, thereby improving procedural justice. This unequal impact on classification F1 score can itself be seen as just under the criterion of prioritarianism, where more resources should be allocated to the individuals or groups that are worse off (Kuppler et al., 2021; Parfit, 2001).

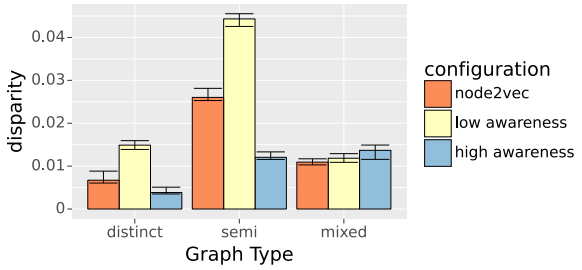


Figure 4. CrossWalk's impact on disparity of sensitive attribute 'location'. The *low awareness* configuration tends to lead to higher disparity, indicating an unequal impact on the different groups.

Finally, we investigate how the *CrossWalk* fairness intervention impacts the classification performance of the non-sensitive attribute. As seen in Figure 5, applying both *low awareness* and *high awareness CrossWalk* configurations leads to decreases in classification performance for the non-sensitive attribute. Furthermore, across all of our conducted experiments, no configuration of *CrossWalk* hyperparameters could improve *performance* over the *node2vec* baseline. This leads to the conclusion that *CrossWalk's* introduces a trade-off between improving fairness for the sensitive attribute and accurately capturing the graph structure.

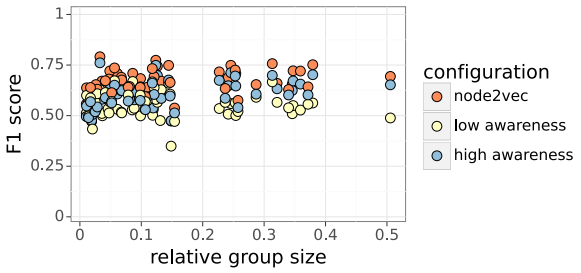


Figure 5. Classification F1 score for the control attribute 'age' by relative group size of groups in all subgraphs. Influencing embeddings for lower or higher awareness of the sensitive attribute makes classifying the control attribute more difficult.

5. Conclusion

Our findings reveal that *CrossWalk* can be strategically parameterized to significantly alter the ability to infer sensitive attributes. This alteration manifests in two principal ways: (1) by enhancing procedural justice through the obfuscation of the sensitive attributes, and (2) by promoting distributive justice via reducing disparities in the classification performance of sensitive attributes.

We demonstrate that these goals are readily achievable: high hyperparameter lead to low awareness of the sensitive attribute, enhancing procedural justice, while low values lead to high awareness of the sensitive attribute, improving distributive justice. Moreover, our results indicate that embeddings generated with *CrossWalk* bias generally outperform those created with node2vec alone regarding these notions of justice. Additionaly, underrepresented groups benefit more from the *CrossWalk* fairness intervention than majority groups.

However, this approach does present a notable trade-off: while it elevates fairness concerning sensitive attributes, it can diminish performance and fairness for other attributes. An important limitation of the *CrossWalk* algorithm is its dependency on prior knowledge of sensitive attributes for biasing. This makes it a powerful tool for deliberately incorporating such information into node embeddings but also presents the potential for misuse. For instance, rather than diminishing the recognition of sensitive attributes, *CrossWalk* could be misused to more effectively identify minority groups, possibly resulting in adverse outcomes for these groups.

Overall, the *CrossWalk* algorithm emerges as a potent tool for learning fairer node embeddings. However, it is important to acknowledge that no single configuration leads to ideal results in every scenario. With the parametrization guidelines presented in this work, *CrossWalk* can be effectively tailored to provide use-case dependent solutions, adeptly navigating the various trade-offs it presents.

Impact Statement

In this paper, we demonstrate how the parametrization of the *CrossWalk* algorithm influences the ability to infer a sensitive attributes from node embeddings. By fine-tuning hyperparameters, we show that it is possible to either significantly enhance or obscure the detectability of these attributes. This functionality offers a valuable tool for improving the fairness of ML systems that rely on graph embeddings, making them adaptable to different fairness paradigms.

However, this technology also raises concerns about potential misuse. Enhanced detection of underrepresented groups, such as those based on religion or sexuality, could be exploited, leading to adverse consequences for these individuals. The *CrossWalk* fairness intervention requires the presence of sensitive attribute values in the dataset, implying that malicious actors already have access to information that could harm underrepresented groups. While the ability to manipulate the prominence of sensitive attributes in embeddings does not fundamentally alter this risk, it underscores the need for responsible use and robust safeguards to prevent misuse.

References

(AI HLEG) High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI. Text, European Commission, April 2019. URL https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Angwin, J. and Parris, T. J. Facebook Lets Advertisers Exclude Users by Race, October 2016. URL https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine Bias, May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=l0i8JndZRzf9U7hmG1DlFV6RjLJo1zYf.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016. Publisher: HeinOnline.

Colquitt, J. A. and Rodell, J. B. Measuring Justice and Fairness. In Cropanzano, R. S. and Ambrose, M. L. (eds.), *The Oxford Handbook of Justice in the Workplace*, pp. 0. Oxford University Press, July 2015. ISBN 978-0-19-998141-0. doi: 10.1093/oxfordhb/9780199981410.013.0008. URL https://doi.org/10.1093/oxfordhb/9780199981410.013.0008.

Goldman, B. and Cropanzano, R. "Justice" and "fairness" are not the same thing. *Journal of Organizational Behavior*, 36(2):313–318, 2015. ISSN 1099-1379. doi: 10.1002/job.1956. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/job.1956. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/job.1956.

Grgic-Hlača, N., Zafar, M. B., Gummadi, K. P., and Weller, A. The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In *Symposium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems*, Barcelona, Spain, December 2016. URL https://www.mlandthelaw.org/papers/grgic.pdf.

Grover, A. and Leskovec, J. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864, San Francisco California USA, August 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939754. URL https://dl.acm.org/doi/10.1145/2939672.2939754.

Hao, K. Facebook's ad algorithms are still excluding women from seeing jobs, April 2021. URL https://www.technologyreview.com/2021/04/09/1022217/facebook-ad-algorithm-sex-discrimination/.

Kaplan, L., Gerzon, N., Mislove, A., and Sapiezynski, P. Measurement and analysis of implied identity in ad delivery optimization. In *Proceedings of the 22nd ACM Internet Measurement Conference*, pp. 195–209, Nice France, October 2022. ACM. ISBN 978-1-4503-9259-4. doi: 10.1145/3517745.3561450. URL https://dl.acm.org/doi/10.1145/3517745.3561450.

Khajehnejad, A., Khajehnejad, M., Babaei, M., Gummadi, K. P., Weller, A., and Mirzasoleiman, B. CrossWalk: Fairness-Enhanced Node Representation Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):11963–11970, June 2022. ISSN 2374-3468. doi: 10.1609/aaai.v36i11.21454. URL https://ojs.aaai.org/index.php/AAAI/article/view/21454. Number: 11.

Kordzadeh, N. and Ghasemaghaei, M. Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3):388–409, May 2022. ISSN 0960-085X. doi: 10.1080/0960085X.2021.1927212. URL https://doi.org/10.1080/0960085X.2021.1927212. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0960085X.2021.1927212.

Kozyrkow, C. AI Bias: Good intentions can lead to nasty results, October 2023. URL https://archive.is/wLAJQ.

Kuppler, M., Kern, C., Bach, R. L., and Kreuter, F. Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There?, May 2021. URL http://arxiv.org/abs/2105.01441. arXiv:2105.01441 [cs, stat].

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

Leskovec, J. and Krevl, A. SNAP Datasets: Stanford Large Network Dataset Collection, June 2014. URL http://snap.stanford.edu/data.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A Survey on Bias and Fairness in Machine Learning, January 2022. URL http://arxiv.org/abs/1908.09635. arXiv:1908.09635 [cs].

Mello-Klein, C. Facebook's ad delivery algorithm is discriminating based on race, gender and age in photos, Northeastern researchers find, October 2022. URL https://news.northeastern.edu/2022/10/25/facebook-algorithm-discrimination/.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient Estimation of Word Representations in Vector Space, September 2013. URL http://arxiv.org/abs/1301.3781. arXiv:1301.3781 [cs].

Miller, D. Justice. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2023 edition, 2023. URL https://plato.stanford.edu/archives/fall2023/entries/justice/.

Parfit, D. Equality and Priority 1. In *The Notion of Equality*. Routledge, 2001. ISBN 978-1-315-19979-5. Num Pages: 20.

Perozzi, B., Al-Rfou, R., and Skiena, S. DeepWalk: online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710, New York New York USA, August 2014. ACM. ISBN 978-1-4503-2956-9. doi: 10.1145/2623330.2623732. URL https://dl.acm.org/doi/10.1145/2623330.2623732.

Rawls, J. *A theory of justice*. Belknap Press of Harvard University Press, Cambridge, Mass, rev. ed edition, 1999. ISBN 978-0-674-00077-3 978-0-674-00078-0.

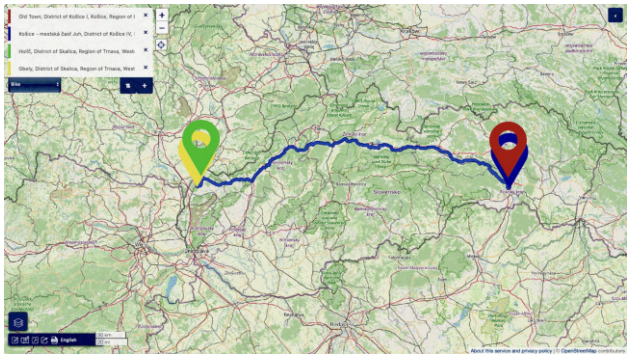
Takac, L. and Zabovsky, M. Data analysis in public social networks. In *International scientific conference and international workshop present day trends of innovations*, volume 1, 2012. URL http://snap.stanford.edu/data/soc-pokec.pdf. Issue: 6.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. LINE: Large-scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pp. 1067–1077, Republic and Canton of Geneva, CHE, May 2015. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-3469-3. doi: 10.1145/2736277.2741093. URL https://dl.acm.org/doi/10.1145/2736277.2741093.

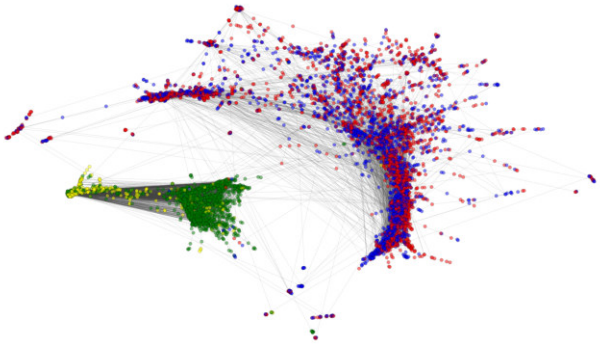
Xu, M. Understanding Graph Embedding Methods and Their Applications. *SIAM Review*, 63(4):825–853, January 2021. ISSN 0036-1445. doi: 10.1137/20M1386062. URL https://epubs.siam.org/doi/10.1137/20M1386062. Publisher: Society for Industrial and Applied Mathematics.

Zhu, X. and Ghahramani, Z. Learning from Labeled and Unlabeled Data with Label Propagation. June 2002.

A. Geographical- and Force Directed Layout of Subgraph ‘semi-distinct 0’



(a) Geographical layout: Displays four towns, each marked with a unique color, representing their geographical locations as foundation for the graph selection.



(b) Graph visualization: Illustrates the social network of individuals from these towns, with each node representing an individual and its color, representing the location group corresponding to the respective town on the map.

Figure 6. Visualization of a semi-distinct graph from our dataset. The left panel shows the geographical positioning of four towns, while the right panel shows the social network derived from these locations. This figure demonstrated the correlation between geographical proximity and network connectivity, where individuals from from geographically closer towns show higher interconnectivity.