Algorithmic Fairness in Social Networks

Hanshu Yu

ABSTRACT

Two hunderd words Two hunderd words

Introduction

Social network analysis has become one of the major ingredients in social sciences. A social network consists of individuals represented as nodes and their interactions are depicted as links between these nodes. Interactions among social networks can be of various forms, for instance, online and offline friendships between people, collaborations between scholars, or just chats during a career fair, etc. Understanding and discovering structures and patterns in social networks has been a fundamental topic that has received large research attention over the past decades.[1]

The task of community detection thus naturally emerges as we cluster individuals into meaningful groups that are called communities based on their interaction patterns. Communities are often intuitively defined as groups of comparably densely connected nodes, meaning individuals are more likely to interact with others within the same community than outsiders.[2–4] There is a wide spectrum of applications of community detection results in social network analysis, ranging from healthcare services[5], link predictions, recommender systems[6], rumor spreading[7], and behavioral analysis of users in online social platforms[8], etc.

People should be fairly treated in these real-world applications of community detection. Unfortunately, despite the large research attention received and the abundant literature about community detection algorithms and applications over the past decades, algorithmic fairness issues in community detection for social networks have largely been ignored. Several works demonstrate that the detected community memberships can largely affect the results of the downstream tasks.[9] Different community detection results can largely influence the link prediction results.[10] Besides, some popular algorithms fail to assign less-connected individuals to proper communities, resulting in such individuals being ignored in subsequent analysis.[11] While the algorithmic fairness of many downstream tasks was evaluated, the algorithmic fairness of community detection algorithms has not been explored. A context-independent fairness evaluation framework will raise the awareness of fairness for users of community detection algorithms.

Here in this paper we carefully discuss the algorithmic fairness of community detection algorithms. We highlight the issues with fairness in community detection problems. A novel individual fairness evaluation framework tailored to community detection is established without relying on the outcome of downstream tasks. The fairness and preference of popular community detection algorithms are examined using an existing synthetic network benchmark model.

Building up the evaluation framework for algorithmic fairness in community detection

Algorithmic fairness has been defined and extensively studied for machine learning classification problems. A commonly accepted framework splits algorithmic fairness into two perspectives: group fairness and individual fairness. Group fairness emphasizes that groups defined by demographic attributes like gender, race, ethnicity, age, etc, should receive equal treatment and resources while individual fairness requires that similar individuals are treated similarly.[9,12,13]

The uniqueness of community detection problem

Over the years many community detection algorithms have been proposed to discover various meaningful community structures.[14,15] This ambiguous definition leaves much flexibility in interpreting how to quantify dense connections and what is

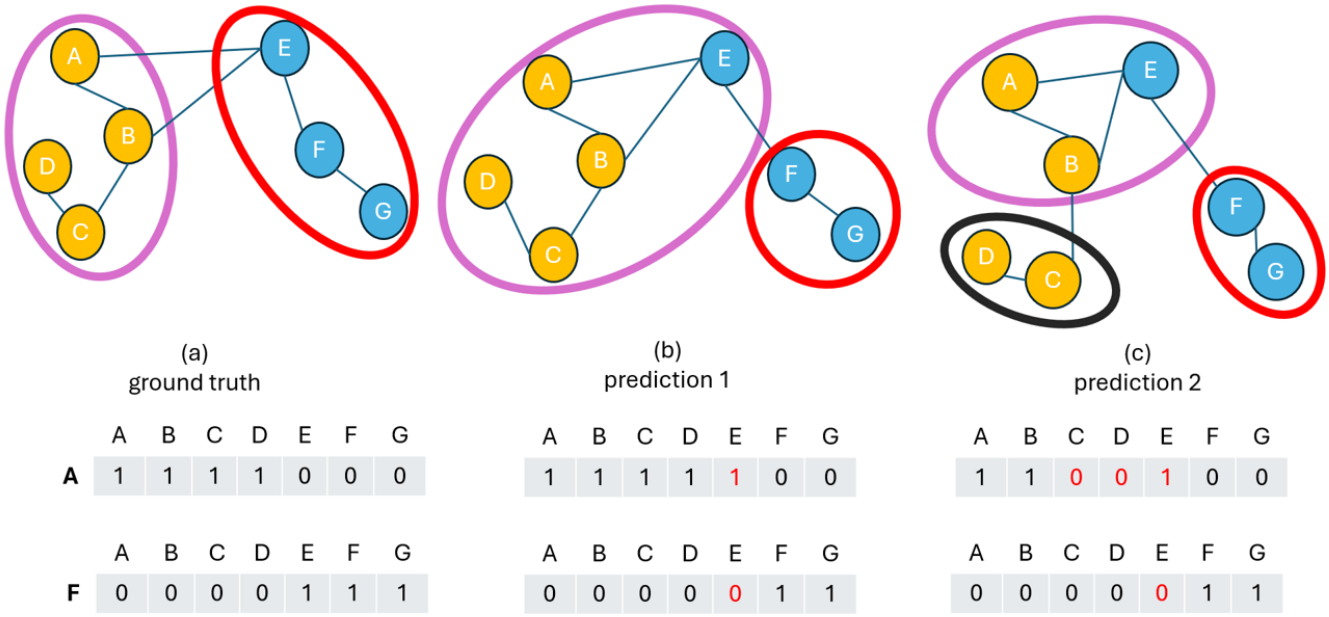

Figure 1. An example of community detection in a toy social network. This toy social network has a ground truth partition defined by sensitive attributes displayed by different nodal colors. The network visualizations on top show different ways of partitioning the network. Mismatches between ground truth and detected structural communities can be clearly observed. In the meantime, community partitions in both (b) and (c) are reasonable partitions satisfying the intuitive definition of communities. The binary arrays at the bottom display the pairwise community membership with the other nodes for node A and node F. 1 indicates that the two nodes are in the same community and 0 otherwise. The red colored numbers in the arrays indicate the disagreements between the prediction and the ground truth. When we deem that for nodes A and F, a fair community detection algorithm should produce the same disagreements compared to the ground truth, then prediction 1 is clearly the fair one compared to prediction 2.

considered a good community. Meanwhile, the meaningfulness of the discovered communities relies on the specific interpretation and the subsequent usage. For example, given a scientific collaboration network, a science news company might seek to discover communities based on the research fields or affiliations to provide better-tailored daily news feeds to the subscribed researchers. However, one might also be curious about discovering personal friendships from co-authorship between researchers. The desired communities might differ greatly in the two cases described before, even though they are detected from the same network. Thus, the precise definition of desirable communities in social networks is largely context-dependent.[14,16]

Group fairness evaluation is largely context dependent

In many cases, we can obtain ground-truth communities from the metadata that represent the desired true results for community detection algorithms to match. However, one important phenomenon is that the ground-truth communities and structural communities can mismatch in the network, not only in terms of disagreements in nodal membership but also disagreements in the number of communities in the network.[10] But we are not claiming that either way of partitioning the network is wrong, we stress the point that the ground truth is not always the whole truth. The ground truth should be treated as a crucial reference point for examining and evaluating community detection algorithms.[17] In general, no algorithm is universally optimal for community detection tasks with respect to detecting ground-truth communities.[?,3] Figure 1 provides a simple and intuitive visualization on the disagreement between ground truth communities and different possible structural communities.

This flexibility in explaining community partition causes trouble when employing group fairness criteria. In real-world applications, the detection result of nodal membership without any downstream information does not directly determine the treatment or resource one might receive. The decision maker, be it an algorithm from a downstream task or a human, is the one who directly determines the allocation based on the nodal membership. For example, if a resource is decided only granted to *community A* out of the detected communities. Although being classified into *community A* seems to grant you the resource

directly, the inequality is caused by the decision maker that makes the unequal allocation criteria but not the community membership information itself. When community membership directly obtained from the community detection algorithm is untied with the context, the sole community membership information does not immediately suggest any kind of unfairness.

Another unresolved issue from the perspective of group fairness is the inconsistency of the number of prediction communities compared to the ground truth. Commonly fairness criteria for classification evaluation are based on the elements obtained in the confusion matrix[18,19], but that poses problems when comparing prediction results based on ground truth partitions. To construct the confusion matrix, a method of mapping the detected communities to the ground truth is required, but that automatically involves a downstream task of mapping prediction to reality in the evaluation. Meanwhile, the mapping method could cause unfairness that requires careful evaluation. While not relying on the confusion matrix, context-dependent scores like distance to centroids or representative members[20] or the loss function in the specific model[21], in clustering are often used as the basis of evaluation.[18,22] But that results in the loss of generality for a group fairness framework.

Because of the issues induced by the disagreement of community partitions and the context-dependent flexibility of interpreting community membership results, a generalized fairness evaluation framework for community detection is largely incompatible from a group fairness perspective. Thus, for the rest of this paper, we focus on the individual fairness of community detection algorithms.

Defining individual fairness criteria

What individual fairness criteria might we infer when we only consider the community detection task? A widely accepted individual fairness criteria requires that similar individuals get similar treatments.[13] However, a reliable nodal similarity measure is very hard to obtain. Even if a similarity measure is present, it is very difficult to infer how this similarity is related to the community membership.[23] If a biased similarity measure is selected, we are at the risk of ignorance: to call an algorithm unfair based on an unfair criterion. Research on individual fairness shows that it is sometimes impossible to construct a similarity metric that accurately reflects task-specific moral values.[24]

Thus, the basis of our fairness criterion should utilize a common equality assumption that all individuals are equal.[19] Our fairness criteria now becomes:

Every individual should receive similar treatment.

We incorporate the merit of the individual fairness definition in fair representation learning that the prediction should be equally consistent for all individuals such that they receive equal treatment from the algorithm.[25] The consistency measured in representation learning is the consistency of classification results obtained for each data item with its nearest neighbors as reference. If some data disagrees with its local neighbors while others agree, then the representation learning algorithm is deemed to be unfair.

In community detection, the consistency is measured using nodal community relationships with all other nodes in the network. For one arbitrary node i in a network with N nodes, the relationship can be represented in a binary vector $\mathbf{y}$ of length N. For each member $\mathbf{y}_j$, $\mathbf{y}_j = 1$ means that node i and j belong to the same community and $\mathbf{y}_j = 0$ otherwise. The fairness criterion requires that the disagreements between the predicted $\hat{\mathbf{y}}$ and the ground truth $\mathbf{y}$ should be equally consistent for all nodes in networks. The different prediction results for nodes A and F in Fig. 1 demonstrate this individual fairness concept. In prediction result 1, both nodal community relationship vectors have only one disagreement with the ground truth. But in prediction result 2, node A has 2 more disagreements than node F. Based on our criteria, the algorithm that produces prediction result 1 is fair while the algorithm that produces prediction result 2 is unfair for nodes A and F.

Common pitfalls of performance leveling-down can occur when incorporating the equality assumption at the individual level.[12] Universal rejection and high risk in common classification problems when everyone is treated equally badly by making wrong predictions or withholding the resource to everyone.[24] In community detection, this levelling-down effect indicates that the predicted relationship vectors $\hat{\mathbf{y}}$ are very different from the ground truth relationship vector $\mathbf{y}$. However, this levelling-down effect in community detection is mitigated by the ambiguity in community definition and the mismatch between ground truth and structural communities. In this case, we do not consider predictions far away from the ground truth memberships to be bad because the consistency in prediction indicates that the existing ground truth might not be the desired community that this algorithm intends to cover. If one algorithm tends to recover the ground truth for some individuals but not for others, we deem the algorithm is treating individuals inconsistently, thus our fairness criterion is violated.

We do not entirely throw the similarity away in the analysis, as they also provide valuable insights into what might the community detection prefer. As we have discovered in related research that some community detection methods treat lowly-connected nodes badly[11], whether an algorithm will unequally treat some node with certain properties inconsistently is of the interest to practitioners that might employ community detection results in their downstream tasks. However, we are cautious to call this preference by the algorithm as unfairness because of the lack of usage context.

Measuring the equality of consistent community membership prediction at individual level

Following the analysis and justification of using the fairness criterion: *every individual should receive similar treatment* by the community detection algorithm. We can formulate a clear and specific fairness criterion for the evaluation of community detection:

Equal treatment*: The individual fairness of a community detection algorithm is achieved by the equality of consistency in community membership prediction at the individual level.*

To evaluate the algorithm, we need a community detection algorithm $\mathscr{A}$, a network G with N nodes, and a N by N ground-truth pair-wise community membership matrix $\mathscr{Y}$ as inputs. Elements $\mathscr{Y}_{ij}$ in $\mathscr{Y}$ take binary values that $\mathscr{Y}_{ij} = 1$ means nodes i and j are in the same community, $\mathscr{Y}_{ij} = 0$ means nodes i and j are in different communities. Such that the roles of $\mathscr{Y}$ are the ground truth relationship vector $\mathbf{y}$ we defined above for all nodes.

Similar to the ground truth matrix $\mathscr{Y}$, the prediction results of a community detection algorithm $\mathscr{A}$ on a network G can be represented in a N by N predicted membership matrix $\hat{\mathscr{Y}}$. Elements $\hat{\mathscr{Y}}_{ij}$ in $\hat{\mathscr{Y}}$ also take on binary values where 1 indicates that nodes i and j are predicted to be in the same communities and 0 otherwise.

For simplicity and completeness, we let $\mathscr{Y}_{ii} = 1$ and $\hat{\mathscr{Y}}_{ii} = 1$.

Because we should measure disagreements between the predicted $\hat{\mathbf{y}}$ and the ground truth $\mathbf{y}$ for each node, we employ a vector distance measure $\mathscr{D}$ to measure the disagreements. The assumption that we incorporate in this measure is the indistinguishability of pair-wise prediction errors at the individual level. This means that for node pair i and j, the error of falsely predicting them into the same community and the error of falsely predicting them into different communities should contribute the same in the distance measure $\mathscr{D}(\mathbf{y}, \hat{\mathbf{y}})$. The single prediction disagreements between different node pairs should also contribute the same in the distance measure $\mathscr{D}(\mathbf{y}, \hat{\mathbf{y}})$.

A direct yet effective distance measure that we use is the Hamming distance. Hamming distance measures the disagreements in two binary arrays. If $\mathbf{y} = [0,1,1,1,0]$ and $\hat{\mathbf{y}} = [0,0,1,1,1]$, then the Hamming distance $\mathscr{D}^{\mathscr{H}}$ between $\mathbf{y}$ and $\hat{\mathbf{y}}$ is 2.

By computing $\mathscr{D}(\mathbf{y}^{(i)}, \hat{\mathbf{y}}^{(i)})$ for each node i in the network, we may obtain a quality vector $\mathbf{q}$ of length N that describes how much disagreement between the predicted nodal membership vector with the ground truth. Each element q_i in $\mathbf{q}$ is calculated as follows:

$$q_i = \frac{\mathscr{D}^{\mathscr{H}}(\mathbf{y}^{(i)}, \hat{\mathbf{y}}^{(i)})}{N-1}$$

For a node i, q_i represents the percentage of disagreeing pair-wise community memberships out of all predictions.

Our defined individual fairness criterion suggests that all values in $\mathbf{q}$ should be equal. When the predicted communities exactly match the ground truth, then $\mathbf{q}$ is a vector of N zeros. In the case of inequality of treatment, the values in $\mathbf{q}$ are different. The more widespread the values are, the more unequal the treatments are, thus the more unfair the community detection algorithm $\mathscr{A}$. We can measure the inequality in $\mathbf{q}$ by calculating the variance in $\mathbf{q}$. A larger variance indicates a larger degree of unfairness exists in the consistency of prediction quality among individuals.

Following the famous Popoviciu's inequality on variances:

$$0 \leq Var[\mathbf{q}] \leq \frac{(M-m)^2}{4}$$

Our inequality measure is bounded by the squared difference between the upper and lower bounds of $\mathbf{q}$. In our case, this is simple, the lower bound of $\mathbf{q}$ is 0 corresponds to the exact match between prediction and ground truth. The upper bound of $\mathbf{q}$ is 1, which corresponds to the complete mismatch between the prediction and ground truth. We can obtain a handy fairness index f:

$$f = 1 - 4\dot{V}ar[\mathbf{q}]$$

The fairness index f is bounded between 0 and 1 with 1 indicating equality of consistency among the individuals. The smaller f is, the more unfair the community detection algorithm's predictions are.

It is easy to spot that an additional benefit of using the Hamming distance is in line with the accuracy equality from the perspective of group fairness where each group should have equal prediction accuracy and the the harm of misclassification should be the same.[12] In our case, if our individual fairness criterion is met, then equal prediction accuracy is automatically achieved. But our individual fairness captures the inequality at the individual level when aggregated prediction accuracy achieves equality for groups, there might still exist large inequality among different groups. However, the equality of aggregated prediction accuracy also does not escape the issue of mapping predicted communities and the ground truth for group fairness.

Evaluating individual fairness using a synthetic benchmark network

We investigate the individual fairness of 20 popular community detection methods using a widely deployed LFR synthetic benchmark network model[26]. LFR model attempts to simulate the scale-free nature of real-world social networks. We can control the fraction of inter-communities links for individuals by varying a parameter μ. A smaller μ indicates that the communities are better separated from each other. The distribution of node degree and community size can also be tuned by two parameters τ_1 and τ_2 in a network. The degree of a node indicates how many links this person has with other individuals. A larger τ_1 means the degree distribution is more skewed while a larger τ_2 means the community size distribution is more skewed.

We test the algorithms in 4 scenarios. The baseline scenario consists of 2500 nodes and well-separated communities with the degree and community size distribution having moderate variations, we test the community detection algorithms and measure their individual fairness index f. In each of the three remaining scenarios, we test the algorithms when the communities are less separated, the distributions of node degree and community size are more skewed, and the distributions of node degree and community size are less skewed.

References

1. Borgatti, S., Mehra, A., Brass, D. J. & Labianca, G. Network analysis in the social sciences. *Science* **323**, 892–895, DOI: 10.1126/science.1165821 (2009).
2. Barabási, A. L. *Network science*, chap. Communities, 320 – 376 (Cambridge University Press, 2016).
3. Peel, L., Larremore, D. & Clauset, A. The ground truth about metadata and community detection in networks. *Sci. advances* **3**, DOI: 10.1126/sciadv.1602548 (2016). 1608.05878.
4. Yang, J. & Leskovec, J. Defining and evaluating network communities based on ground-truth. *Knowl. Inf. Syst.* **42**, 181–213, DOI: 10.1007/s10115-013-0693-z (2015). 1205.6233.
5. Rostami, M., Oussalah, M., Berahmand, K. & Farrahi, V. Community detection algorithms in healthcareapplications: A systematic review. *IEEE Access* **11** (2023).
6. Gasparetti, F., Sansonetti, G. & Micarelli, A. Community detection in social recommender systems: a survey. *Appl. intelligence* DOI: 10.1007/s10489-020-01962-3 (2020).
7. Gupta, K. & Potika, K. Fake news analysis and graph classification on a covid-19 twitter dataset. DOI: 10.1109/BigDataService52369.2021.00013 (2021).
8. Ali, M., Kifayat, M. H. K., Kim, J. Y., Hakak, S. & and, M. K. K. Social media content classification and community detection using deep learning and graph analytics. *Technol. Forecast. Soc. Chang.* **188** (2023).
9. Saxena, A., Fletcher, G. & Pechenizkiy, M. Fairsna: Algorithmic fairness in social network analysis. *ACM Comput. Surv.* **56**, 1–45, DOI: 10.48550/arXiv.2209.01678 (2024). 2209.01678.
10. Ghasemian, A., Hosseinmardi, H. & Clauset, A. Evaluating overfit and underfit in models of network community structure. *IEEE Transactions on Knowl. Data Eng.* **32**, DOI: 10.1109/TKDE.2019.2911585 (2018). 1802.10582.
11. Mehrabi, N., Morstatter, F., Peng, N. & Galstyan, A. Debiasing community detection: The importance of lowly connected nodes. *nternational Conf. on Adv. Soc. Networks Analysis Min.* DOI: 10.1145/3341161.3342915 (2019). 1903.08136.
12. Mittelstadt, B., Wachter, S. & Russell, C. The unfairness of fair machine learning: Levelling down and strict egalitarianism by default. *Mich. Technol. Law Rev.* DOI: 10.48550/arXiv.2302.02404 (2023). 2302.02404.
13. Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. Fairness through awareness. 214–226, DOI: 10.1145/2090236.2090255 (2012). 1104.3913.
14. Fortunato, S. & Newman, M. 20 years of network community detection. *Nat. Phys.* **18**, 848–850, DOI: 10.1038/s41567-022-01716-7 (2022). 2208.00111.
15. Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Phys. Reports* **659**, 1–44, DOI: 10.1016/j.physrep.2016.09.002 (2016). 1608.00163.
16. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. Defining and identifying communities in networks. *Proc. national academy sciences* **101**, DOI: 10.1073/pnas.0400054101 (2004). cond-mat/0309488.
17. Hric, D., Darst, R. & Fortunato, S. Community detection in networks: Structural communities versus ground truth. *Phys. review. E* **90**, DOI: 10.1103/PhysRevE.90.062805 (2014). 1406.0146.
18. Caton, S. & Haas, C. Fairness in machine learning: A survey. *ACM Comput. Surv.* **56**, 1–38, DOI: 10.1145/3616865 (2024). 2010.04053.

19. Mitchell, S., Potash, E., Barocas, S., D'Amour, A. & Lum, K. Algorithmic fairness: Choices, assumptions, and definitions. *Annu. Rev. Stat. Its Appl.* DOI: 10.1146/annurev-statistics-042720-125902 (2021).

20. Abbasi, M., Bhaskara, A. & Venkatasubramanian, S. Fair clustering via equitable group representations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 504–514, DOI: 10.1145/3442188.3445913 (Association for Computing Machinery, New York, NY, USA, 2021).

21. Makarychev, Y. & Vakilian, A. Approximation algorithms for socially fair clustering. DOI: 10.48550/ARXIV.2103.02512 (2021). 2103.02512.

22. Buet-Golfouse, F. & Utyagulov, I. Towards fair unsupervised learning. *Conf. on Fairness, Accountability Transpar.* DOI: 10.1145/3531146.3533197 (2022).

23. Castelnovo, A. *et al.* A clarification of the nuances in the fairness metrics landscape. *Sci. Reports* DOI: 10.1038/s41598-022-07939-1 (2021). 2106.00467.

24. Fleisher, W. What's fair about individual fairness? *AAAI/ACM Conf. on AI, Ethics, Soc.* DOI: 10.1145/3461702.3462621 (2021).

25. Zemel, R., Wu, L. Y., Swersky, K., Pitassi, T. & Dwork, C. Learning fair representations (2013).

26. Lancichinetti, A., Fortunato, S. & Radicchi, F. Benchmark graphs for testing community detection algorithms. *Phys. review E* DOI: 10.1103/PhysRevE.78.046110 (2008). 0805.4770.

LaTeX formats citations and references automatically using the bibliography records in your .bib file, which you can edit via the project menu. Use the cite command for an inline citation, e.g.[?].

For data citations of datasets uploaded to e.g. *figshare*, please use the `howpublished` option in the bib entry to specify the platform and the link, as in the `Hao:gidmaps:2014` example in the sample bibliography file.

Acknowledgements (not compulsory)

Acknowledgements should be brief, and should not include thanks to anonymous referees and editors, or effusive comments. Grant or contribution numbers may be acknowledged.

Author contributions statement

Must include all authors, identified by initials, for example: A.A. conceived the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the manuscript.

Additional information

To include, in this order: **Accession codes** (where applicable); **Competing interests** (mandatory statement).

The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.