Link Prediction and its Ethical Implications

Abhinav Agnihotry 15667766, Brent Brakenhoff 13068989,
Jonathan van der Maas 15248844, Shantanu Motiani 13757830

Group A3

1 Technical Solution

1.1 Abstract

This two-part project proposes and evaluates a link prediction model to predict missing edges in a network and contains an essay questioning the ethics of link prediction in an e-commerce environment. The link prediction model was trained on a combination of graph based features, such as common neighbours between nodes, the Jaccard coefficient and the Adamic-Amar index. Multiple methods were evaluated and logistic regression resulted in the highest accuracy scores. The model can predict the if an edge between two nodes exists with an accuracy of 0.8226. The model ended with a demographic parity difference of 0.0499 and an equalised odds difference of 0.1100, meaning there is a low difference of positive prediction rates between groups, but there is a slight significant difference between the difference of true and false positive rates between node attributes.

The essay delves into ethical concerns, particularly fairness, highlighting how historical data biases and behavioral patterns may disproportionately impact marginalized groups like low-income consumers and emerging brands. Proposed mitigations, including bias audits and fairness constraints, aim to reduce these effects. While the essay mainly addresses fairness, it acknowledges the intertwined issues of privacy and transparency, suggesting the need for future research into their impact and the practical implementation of fairness measures at scale. Addressing these concerns can contribute to a more inclusive and equitable online marketplace.

1.2 Introduction

By using machine learning techniques and graph-based feature engineering, this project aims to build a model that can predict missing edges in a network based on structural features of the graph and a single attribute assigned to each node. The network, pictured in figure 1 consists of 1500 nodes and 6600 edges, however from this network 733 edges were intentionally removed. The proposed model will be required to accurately identify which edges originally existed in the graph, from a set of 1466 candidate edges.

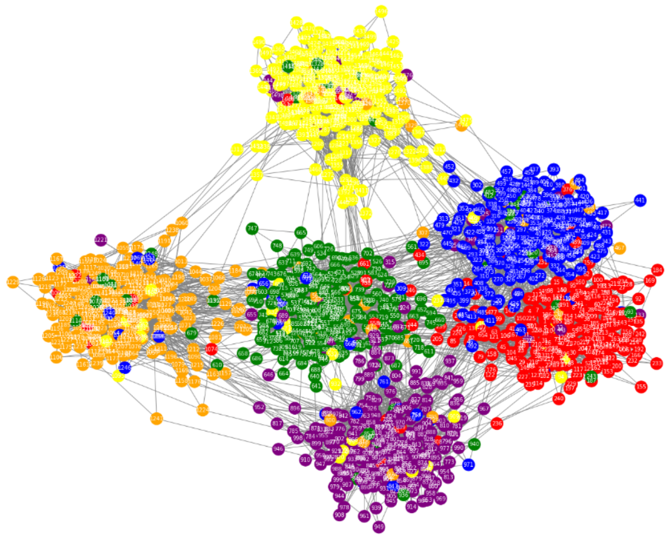

Figure 1: The network

1.3 Feature engineering

Each node in the network is labelled by one of 5 different attributes: ‘x’, ‘d’, ‘y’, ‘f’, ‘m’ and ‘l’. These labels were used to derive one of the features used in the solution to this problem. The simplest usage of these attributes is creating a binary variable, which is 1 if the two nodes in question possess the same attribute and 0 otherwise. Given that there is an edge, there is a 71% chance that these nodes have the same attribute.

However to account for the fact there also exist edges between nodes that have different attributes, one-hot encoding was utilised. Every attribute is represented by a unique vector. For two nodes these two vectors where appended, resulting in a unique vector for every unique combination of attributes.

Other common features for link prediction algorithms were also taken into consideration, taken from the NetworkX documentation. Such as the preferential attachment, which is the number of neighbours of one node multiplied by the number of neighbours of the other; the amount of common neighbors; the Jaccard coefficient, which is the number of common neighbours of two nodes divided by the combined amount of neighbours the two nodes have; the Adamic-Adar index, which returns $\frac{1}{Log(len(\Gamma w))}$ where Γw is the intersection of the sets of neighbours of both nodes; and the resource allocation index, which is similar to the Adamic-Adar index, but without taking the Log() of the length of the intersection.

These features were all used in the final model, favouring the simple binary variable over One-hot encoding, as One-hot encoding did not end up improving the accuracy and increased computational complexity. The shortest distance between nodes was examined briefly but disregarded as the graph is not transitive and thus it caused an unrealistically high accuracy.

1.4 Data preparation

In order to do proper feature engineering, first the data had to be prepared for it. The initial data consisted of two files, one containing 6600 out of the 7333 edges of the original network and one csv file with the attribute associated with each node. First all attributes were assigned to their respective nodes by combining the two files which resulted in 1. This figure shows the network with all provided edges and the attribute for each node where each colour represents one of the different possible values.

After having added all the attributes to their respective nodes, negative edges had to be sampled in order to create a balanced dataset. This was done by randomly selecting pairs of nodes that do not have an existing edge between them, ensuring that the number of sampled negative edges matched the number of positive edges in the graph. This process maintains a 50/50 distribution between positive and negative examples, which was crucial for training our model without bias towards one class over the other. Negative samples were carefully selected to avoid self-loops and duplicate edges.

As a final step in the data preparation, our data, including the new negative samples, was split in a training and testing set before fitting different models to the data to see which performed the best.

1.5 Model selection

To predict the existence of links between nodes in the graph, several machine learning models were evaluated, including Logistic Regression, Random Forest, and K-Nearest Neighbors. Each model was assessed based on its ability to handle the complexities of graph data and produce reliable classifications of both positive and negative edges.

The first model used was Logistic Regression, which served as the baseline for comparison. Logistic Regression was chosen for its simplicity and interpretability, providing an initial assessment of how well linear decision boundaries could capture the relationships between node pairs. Hyperparameter tuning, such as adjusting regularization strength and iterations, was applied to optimize the model's performance and reduce the risk of overfitting.

Following this, Random Forest and K-Nearest Neighbors (KNN) were considered to address potential non-linear relationships within the data. Random Forest, an ensemble method, offered the potential to capture more complex interactions by aggregating multiple decision trees. KNN was explored as a distance-based approach, where the classification of a node pair was based on the majority class among its closest neighbors in feature space. Despite their flexibility, these models were ultimately found to add unnecessary complexity without substantial performance gains. This, along with its interpretability and ease of implementation, made Logistic Regression the most suitable model choice.

1.6 Model training and validation

The models were trained and validated on the graph in figure 1. As the list of output data contained an equal amount of real edges as fabricated edges with no indication which ones are which, this could not be used as testing data. At this point the data consists of two arrays, an array X consisting of the list of features corresponding to each edge, and an array Y consisting of the label for each edge, where 1 means the edge does exist in the graph and 0 means it does not. These arrays were split into a training set and a test set using a 70/30 ratio, this ratio was chosen so that there would be enough data for the model to be trained, while the test set would contain enough data for an accurate evaluation of the model. Finally Grid search was used to fine-tune the hyperparameters of the model.

1.7 Final results

The model which, during model selection, was found to result in the highest cross-validation accuracy was the logistic regression model. This model, including the features mentioned in the sections above, initially resulted in a mean cross-validation accuracy of 0.8212. It was, however, crucial for further fairness analysis that the sensitive attribute itself was also taken into account during the feature engineering. This slightly reduced the mean cross-validation accuracy to 0.8108. Attempting to increase the accuracy by hyperparameter tuning using GridSearchCV barely made a difference as the accuracy increased to 0.8124. Plotting the confusion matrix of the model resulted in the following metrics:

- **Accuracy**: 0.8226
- **Precision**: 0.8216
- **Recall**: 0.8294
- **F1 Score**: 0.8255

Based on the metrics above, it can be concluded that there is room for improvement when it comes to accuracy but initially balancing the dataset did result in a similar precision and recall resulting in a balanced model. The confusion matrix itself can be found in the appendix at figure 2.

As can be seen in table 1, the different metrics from the confusion matrix was also plotted for the different attributes together with the TPR and FPR per sensitive group in table **??**. Across the different groups, very similar metrics were achieved with group x standing out with lower accuracy, precision and recall and also a slightly higher false positive rate. Altogether this resulted in a demographic parity difference of 0.0499 and an equalised odds difference of 0.1100. This shows that, although the rate of positive predictions between groups is very similar, the difference in true and false positive rates between the different groups is slightly high.

Group	**Accuracy**	**Precision**	**Recall**	**F1 Score**
d	0.8449	0.8149	0.8835	0.8478
f	0.8533	0.8489	0.8567	0.8528
l	0.8451	0.8447	0.8659	0.8552
m	0.8281	0.8424	0.8152	0.8286
x	0.7624	0.7837	0.7541	0.7686
y	0.8052	0.7934	0.8094	0.8013

Table 1: Confusion Matrix Metrics per Sensitive Group

Sensitive Feature	True Positive Rate	False Positive Rate	False Negative Rate
d	0.857605	0.201923	0.142395
f	0.857988	0.157143	0.142012
l	0.865385	0.139159	0.134615
m	0.811047	0.201807	0.188953
x	0.755376	0.195652	0.244624
y	0.798658	0.181529	0.201342

Table 2: Fairness metrics by sensitive feature

1.8 Limitations

This approach might not generalise to larger graphs with more node attributes. Logistic Regression worked well in this case, but might not be optimal for real world cases due to the simplicity of the algorithm. Some of the graph features that were used to train the model are similar, such as the resource allocation index and the Adamic-Adar index, however this did not affect model accuracy and only has minimal computational effects. The project also did not consider Graph Neural Networks, which may have resulted in a higher accuracy, as those were not part of the course curriculum.

2 Essay:

Ethical Implications of Link Prediction in e-commerce websites: A Focus on Fairness

2.1 Introduction

Link prediction plays a crucial role in various applications, including online shopping, where it is used to recommend products, personalise offers, and predict future consumer behaviour (Waikhom & Patgiri, 2023). We use logistic regression to perform link prediction, benefiting from its simplicity and interpret-ability. However, using link prediction in e-commerce, or any other context, raises several ethical concerns, specifically regarding Fairness, Privacy, and Transparency.

Fairness: Link prediction models may reinforce biased product recommendations, leading to unequal treatment of customers and limit the reach for the upcoming brands. Customers from certain demographics, such as those with lower incomes or less purchasing power, may receive fewer or less relevant product recommendations compared to higher-income or frequent shoppers.

Privacy: Link prediction models rely heavily on personal data, including purchase history, browsing habits, and customer interactions, to generate recommendations. This extensive data usage can lead to privacy concerns, as the model may inadvertently expose sensitive information about customers' preferences, behaviours, or lifestyle choices through personalised offers.

Transparency: Although logistic regression is relatively transparent, the decision-making behind recommendations may still be unclear to customers, raising concerns about algorithmic transparency. This can undermine trust, as users might suspect recommendations are driven by profit maximisation rather than fairness, leading to accountability concerns.

This essay focuses on Fairness, examining its impact on consumers and newer brands in the context of product recommendations and personalised offers on online shopping platforms. It explores fairness issues from both consumer and emerging brand perspectives, and proposes potential solutions to address these challenges.

2.2 Elaboration

Link prediction in online shopping platforms like Amazon and Walmart relies on customer behaviour data to generate product recommendations and personalised offers. Our Logistic regression model if trained on purchase history, browsing patterns, and interaction data, can predict which products are likely to be of interest to consumers. However, this can lead to fairness concerns, particularly for marginalised customer groups and newer brands that struggle to compete with more established brands for visibility.

2.2.1 Potential Causes:

Bias in training data: The product recommendation system is trained on historical customer data, which may reflect pre-existing biases toward specific demographics, such as high-income shoppers or frequent purchasers. As a result, the model may predominantly recommend products to these favoured groups while offering limited options to lower-income or occasional shoppers.

Reinforcement of brand dominance: Link prediction models depend on user behaviour, such as purchase frequency and product reviews. Customers who frequently buy from established brands reinforce the model's bias toward those brands. This creates a group-fairness issue, as newer brands,

which lack substantial historical data, struggle to achieve visibility in personalised recommendations. Consequently, they face significant challenges in penetrating the market, perpetuating the dominance of well-known brands and limiting consumer choice.
Behavioural biases: Link prediction models rely on customer behaviours like purchase frequency, cart additions, and product reviews. Customers who engage less frequently may receive fewer or less relevant recommendations, which reinforces existing inequalities. This scenario illustrates emergent bias, as the system's operation leads to sidelining less active users and ultimately discouraging their engagement with the platform.

2.2.2 Affected demographic and the potential impact

Low-income or infrequent shoppers *Potential Impact: Deepening Consumer Inequity* Low-income or infrequent shoppers are often excluded from personalised offers and product recommendations, as algorithms prioritise frequent or higher-spending customers. This results in fewer deals for these groups, creating a feedback loop that limits access to potential savings and perpetuates economic disparities. As wealthier customers benefit from tailored offers, the gap widens, reinforcing inequality. This is an example of emergent bias, where unintended inequalities arise from user interactions. It parallels the predictive policing example from the lecture, but instead of increased police presence, frequent buyers receive more offers.

Underrepresented demographic groups *Potential Impact: Market segmentation* When the training data for algorithms is unrepresentative, underrepresented demographic groups—often from diverse geographic, socioeconomic, or cultural backgrounds—receive less relevant recommendations. This lack of inclusively leads to market segmentation, where wealthier customers see premium products while underrepresented groups are directed toward more affordable options. Consequently, this reinforces existing divides, limiting product variety for underrepresented consumers and entrenching inequalities. This situation parallels the findings in (Zhao, Wang, Yatskar, Ordonez, & Chang, 2017), which talks about gender biases on the shopping platforms.

Newer or smaller brands *Potential Impact: Barrier to market entry for new brands* Emerging brands or those with limited historical data struggle to compete against established brands that dominate recommendation systems (Hoogeveen, 2023). This lack of visibility hampers their ability to gain traction, as customers are consistently presented with products from familiar brands. Consequently, established brands receive repeated recommendations due to their historical engagement, while newer brands are sidelined, stifling innovation and competition. Over time, this reinforces consumer behaviour and limits product diversity, creating a feedback loop that discourages exploration of new or niche offerings.

E-commerce brands use link prediction algorithms for product recommendations, but these systems often favour products based on spending habits, resulting in unequal offers. Lower-income consumers receive fewer discounts, worsening online shopping inequality, while newer brands struggle to compete as established products dominate. Consequently, both customers and businesses suffer from an unfair recommendation system.

In the context of the link prediction model, we can assess fairness by examining the True Positive Rate (TPR), False Positive Rate (FPR), and False Negative Rate (FNR) across different groups, as shown in Table 1. Groups like "x" and "y" exhibit higher False Negative Rates (24.46% and 20.13%, respectively), which means they are less likely to receive positive outcomes when they should. Similarly, groups "m," "x," and "d" face higher False Positive Rates (20.18%, 19.57%, and 20.19%, respectively), indicating a higher likelihood of incorrect positive predictions.

These disparities highlight concerns around equalized odds, where the prediction model does not ensure equal error rates across groups. While the differences may seem modest, if left unchecked, unequal odds can become amplified over time, reinforcing biases and reducing fair representation for underperforming groups. These underperforming groups can be the infrequent buyers or the smaller brands which are new to the space.

Moreover, the issue of demographic parity arises, where certain groups (like "x" and "y") are systematically less likely to receive favorable outcomes compared to others, which could lead to long-term disadvantages. If the algorithm begins with unequal treatment, it risks perpetuating and deepening these biases, creating a model that is inherently unfair to specific groups.

2.3 Mitigation

To address fairness concerns in link prediction models used for product recommendations, online shopping platforms must adopt several strategies aimed at ensuring equitable treatment of all customers.

2.3.1 Data-Level Mitigation

Bias auditing and data balancing: Regular audits of the customer data used for model training can help identify and correct biases that disproportionately favor certain demographics. Ensuring that all customer groups are equally represented in the training data, including low-income shoppers and those with non-mainstream preferences, is key to creating a fairer recommendation system.

Data augmentation: In situations where certain customer groups are underrepresented (e.g., niche shoppers or those with less purchasing power), data augmentation techniques can be used to create synthetic data points. This ensures that the model has enough diverse data to learn from and reduces bias caused by data scarcity.

2.3.2 Model-Level Mitigation

Fairness constraints: Constraints can be introduced into the logistic regression model to ensure that recommendations are distributed equitably. For example, demographic parity constraints can ensure that product recommendations are balanced across income levels, geographic regions, and other demographic factors.

Personalization with diversity: While the goal of link prediction is to personalize product recommendations, diversity should also be incorporated into the recommendation engine. This means that, alongside personalized offers, customers should be shown products that they may not have considered based on their previous behavior, encouraging diverse shopping experiences.

Context-aware recommendations: Instead of relying solely on past purchasing behavior, models can incorporate contextual factors such as economic conditions or cultural events that may influence shopping habits. This would allow platforms to offer more contextually relevant recommendations, which are less likely to be influenced by past biases.

2.3.3 Ongoing Monitoring and Auditing

Fairness monitoring metrics: Once deployed, the recommendation system should be continuously monitored using fairness metrics such as disparate impact or equalized odds to ensure that all customer groups are receiving appropriate and equitable recommendations. Any disparities that arise should be addressed through retraining or model adjustments.

Human oversight: In certain high-impact scenarios, such as personalized pricing or exclusive offers, human oversight can provide an additional layer of fairness. Manual review processes can ensure that offers are not disproportionately benefiting one group over another and that all customers have equal access to discounts and promotions.

2.3.4 Customer-Centric Solutions

User feedback integration: By allowing customers to provide feedback on recommendations and offers, platforms can actively involve them in the personalization process. This creates a feedback loop where customers' preferences are continually refined, making the recommendations fairer and more relevant.

Transparency in recommendations: Platforms should aim to make the recommendation process more transparent to consumers, providing explanations for why specific products or offers were recommended. This helps build trust between the platform and its users, ensuring that customers understand the factors influencing their personalized shopping experience.

2.4 Conclusion

This essay examines the ethical concerns related to fairness in link prediction models used for product recommendations. Our analysis highlights how historical biases in training data, behavioral tendencies, and feature selection can disproportionately affect marginalized groups, such as low-income or niche consumers, and limit opportunities for newer brands. We believe that the proposed mitigations, such as bias auditing, fairness constraints, and customer feedback integration, are essential in creating a more equitable system. While our focus has primarily been on fairness, privacy and transparency issues also deserve attention in future work, as these problems can intertwine. One limitation of this essay is that it primarily concentrates on fairness without delving into the technical feasibility or costs associated with implementing the mitigations. Further research is necessary to explore how these proposed solutions can be practically deployed at scale. By addressing fairness systematically, platforms can reduce bias and improve the overall shopping experience, fostering a more inclusive and equitable digital marketplace

References

Hoogeveen, M. (2023). Open data policy, connectivity and portfolio determine brand popularity in e-commerce channel. doi: 10.14293/s2199-1006.1.sor-.pphs6ie.v1

Waikhom, L., & Patgiri, R. (2023). Chapter three - an empirical investigation on biggraph using deep learning. In R. Patgiri, G. C. Deka, & A. Biswas (Eds.), *Principles of big graph: In-depth insight* (Vol. 128, p. 107-133). Elsevier. Retrieved from https://www.sciencedirect.com/science/article/pii/S0065245821000607 doi: https://doi.org/10.1016/bs.adcom.2021.09.007

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017, September). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2979–2989). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D17-1323 doi: 10.18653/v1/D17-1323

3 Appendix

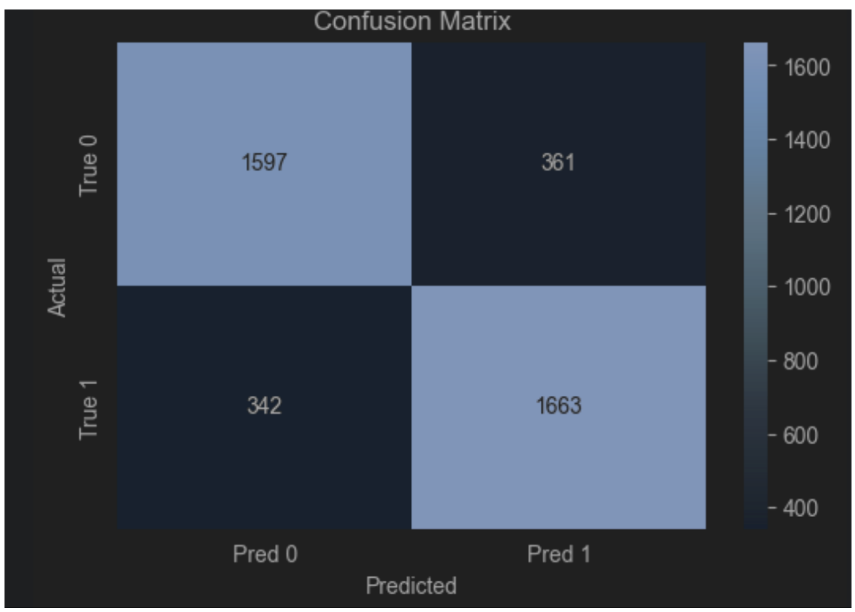

Figure 2: The confusion matrix of the final model