Master’s Program Applied Data Science

Utrecht University 2023/2024

MASTER THESIS

Optimizing Graph Neural Networks for Predicting Fertility Intentions in Social Networks

Dimitrios Diamantidis (7766408)

Supervisors:

Javier Garcia-Bernardo

Date of submission: 01/07/2024

TABLE OF CONTENTS

1. ABSTRACT

This research explores the use of Graph Neural Networks (GNNs) to predict fertility intentions in social networks, to see if GNNs can capture the complex social structures that influence individual decisions. Despite using advanced techniques like iterative model development, targeted optimisation, over-sampling and cross-validation the results are moderate with the highest F1-scores around 50%. This shows the challenges GNNs face when applied to complex and imbalanced data common in social sciences. The results show that GNNs can identify patterns in social networks that affect fertility choices but the accuracy is limited. This research highlights the need for more advancements in machine learning to better handle social network data, GNNs are promising but need a lot of work to be accurate in social science applications.

2. INTRODUCTION

2.1 MOTIVATION AND CONTEXT

Predicting individual decisions in social contexts, especially fertility intentions, is a big challenge and opportunity in social sciences. Traditional statistical models can't capture the subtleties of social networks on such personal decisions. This research is driven by the need to improve models with better handling of complex social data, an area where current methods are lacking.

Stulp et al. (2023) go into great detail about the interplay between individual traits and social network structure on fertility decisions. Their analysis shows that while personal characteristics like age, education and economic status are the main drivers of fertility choices, the social network (friends, family and broader social connections) also plays a big role, but is less studied. They quantitatively show that people are significantly influenced by the fertility patterns and attitudes of those in their social network, and that there is a form of social contagion in fertility intentions. This contagion effect is small but people often align their reproductive choices with the norms and trends they see in their network.

Additionally Stulp et al. (2023) point out that current statistical models often miss these network dynamics, they focus on individual direct factors without accounting for the broader social interactions that are equally important. They argue that this oversight means there is a big gap in our understanding of the full range of influences on fertility decisions, particularly how indirect influences like peer effects and network centrality shape an individual's choices. Their results show we need more advanced analytical tools that can combine complex network data with traditional demographic variables to get a full picture of what drives fertility decisions.

By pointing out these gaps Stulp et al. (2023) not only show the importance of social networks in fertility decisions but also lay the ground for using more advanced computational methods like Graph Neural Networks to model these interactions better. Their work is a foundation for this thesis, and provides the motivation to use GNNs to understand and predict how social structures and individual decisions intersect in fertility.

Unlike traditional models, GNNs are designed to handle data with relational information, like social networks. They can process and learn from both node and edge data, which makes them perfect to untangle the complex interdependencies and subtle social signals that influence fertility decisions. The use of GNNs in this study goes beyond the traditional use, driven by recent advances in neural architecture search (NAS) that can optimize GNNs for specific tasks. As Zhou et al. (2022) show, automated configuration of GNNs can lead to significant performance improvements across different datasets, which is key to adapt models to the diverse and dynamic nature of social data.

2.2 LITERATURE OVERVIEW

We've seen the rise of Graph Neural Networks (GNNs) for complex data structures across many domains and it's clear they can find patterns in big messy data. A survey by Gori et al. (2021) on GNN architectures and optimisations goes over the foundational models that led to the recent innovations and how GNNs can adapt to different data and structures. This foundational work shows how GNNs are particularly good at preserving the relational information that's key to correct data interpretation which is important for social sciences and beyond. More recently Zhang et al. (2022) in Neural Architecture Search (NAS) for GNN-based graph classification show how automated methods can optimise GNNs without human intervention. This shows how NAS can massively improve the efficiency and effectiveness of GNNs, adapt the architecture to the specific dataset. In the context of social networks, Bronstein et al. (2020) on learning combinatorial optimisation on graphs show how GNNs can be used to address scalability and computational efficiency. This is especially important when dealing with big social networks where the data is too big for traditional methods to handle.

Also GNNs in social behaviour is shown by studies like Kipf and Welling (2019) which show how GNNs can predict node properties based on their connections in the network. This is a direct link to the exploration of social influence in networks, how GNNs can model complex interpersonal dynamics that affect individual decisions like fertility intentions. Overall these studies show the technical progress in GNN and the practical applications, so we can use GNN to explore the subtle influences of social networks on individual behaviour. This is the background of this study which will extend these methodologies to the domain of fertility intentions and show the broader implications of GNN in theory and practice in social sciences.

2.3 RESEARCH QUESTION

Based on the literature review this study moves on to a more detailed exploration of how advanced computational methods can be applied to specific sociological phenomena. The refined research question guiding this thesis is: How can we optimise the prediction ability of Graph Neural Networks in a social science context? How can we use Graph Neural Networks to effectively model and forecast fertility intentions within social networks?

This question highlights the dual challenge of using complex machine learning techniques like GNNs to not only understand but also predict human behaviour based on social interactions. By focusing on optimising the GNN's predictive ability we can get deeper insights into how individual fertility decisions are influenced by the broader network dynamics. This approach combines technical refinement with sociological inquiry to develop models that are not only theoretically sound but also practically applicable to predict and understand subtle social behaviour. The outcomes of this research will have a big impact on policy development, targeted interventions and the field of social network analysis making it a bridge between data science innovation and real world applications.

3. DATA

3.1 SELECTED DATA EXPLORATION RESULTS

Personal or egocentric network data collection is a key way to study the impact of social networks on fertility decisions. In this approach, respondents (egos) list several key people in their network (alters) and provide detailed information about these people and their connections (ties) to them. This is the same method used by Stulp et al. (2023) so I know it works to capture a range of social influences on fertility behaviour.

Participants in this study were asked to list friends and acquaintances (alters), describe the quality of these relationships (tie strength) and provide information about the behaviour and characteristics of these alters. This allows me to look at various social influence processes, such as the impact of tie strength, where close ties with family or peers who have recently had children might encourage pro-natal attitudes. And the composition of an individual's network, which includes the characteristics of the alters, reflects the types and availability of resources in the network, such as support from kin in child-rearing contexts.

An important part of personal network data collection is deciding on the number of alters to include. Larger networks provide more data but put more burden on the respondent in terms of survey length and complexity. Following Stulp et al. (2023) and other studies I decided to limit the number of alters to 25 to balance detail with respondent burden. This is considered enough to capture network composition and structure without overwhelming participants.

During the data exploration stage, we found some interesting things. The age range was 18 to 50 and 30 was the most common age (4% of all records). That is a wide age range and allows us to look at fertility intentions across different life stages. The number of children respondents had varied from 0 to more than 5 and most of the alters had two, over 19%. That is good to see, as it gives us insight into how many children someone already has and how that affects future fertility intentions. Many respondents had strong social connections, both within and outside of family structures according to network analysis. This included spending a lot of time with close friends and family who were thought to have a social influence and supportive role on people's intentions to have a baby.

Looking at the different types of connections and interaction frequencies in the survey was part of the data exploration process. Turns out face to face contacts are more common among broader social ties, 25% of the total alters have face to face interactions.

These exploratory results informed the next stages of the study especially when it came to adjusting the GNN architecture to account for what was found and tuning the feature selection process to focus on the variables that mattered most. The results of this phase provided a good foundation to build an advanced analytical approach that would capture the complex relationships between network and individual variables that influence reproductive intentions.

3.2 DATA PREPARATION FOR ANALYSIS INCLUDING MOTIVATION

To make sure the data was accurate and useful for the Graph Neural Network (GNN) used in my study I had to take several precautions before I could prepare the dataset for analysis. To ensure uniform treatment of the data the first step was to standardise the essential columns.

To speed up the examination of the participants' fertility intentions I created a mapping system where "Probably so" and "Absolutely so" were noted as "2" meaning they wanted to have children and "Probably not" and "Absolutely not" were noted as "1" meaning they didn't want to have children and the "i don't know" responses were noted as "0" meaning they were neutral. This made it easier to analyse the data in a more systematic and quantitative way.

This mapping it's more systematic and reduce variability within the response categories. It also minimize subjective interpretation errors that might occur when different researchers or algorithms try to categorize the responses. It also provide a clear and straightforward way to visualize the trends and patterns in the data and to compare between different groups of respondents based on their fertility intentions.

A lookup dictionary was created to speed up data handling and analysis. Because of this, fertility intentions could be added to the graph structure of the GNN, with each participant's data as a node with attributes representing their relationship and personal information. Besides these changes, the data went through advanced preprocessing to fill in missing values and encode categorical variables like gender and relationship types. Categorical variables were transformed to fit the machine learning model and continuous variables were standardised to help the GNN process and learn from the network data.

We handled missing values in the dataset on a variable-by-variable basis. For example, ages listed as ‘50+’ were set to 50 ensuring a consistent upper limit, and ages under 18 were set to 18. Missing ages were set to 0 to make it clear that the data was missing. The same logic applied to the number of children, missing entries and "I don’t know" were set to 0, and "Expecting first child" was set to 0.5 since they were about to become parents. This systematic treatment of missing data allowed our GNN to process and learn from the network data without being confused by the gaps.

Categorical variables were encoded precisely to convert them into a format for the graph neural network analysis. Gender was encoded as binary, with ‘Female’ as 1 and ‘Male’ as -1, so this attribute was in a simple numerical format. Relationship types were also distilled into numerical values, converting descriptive relationship data into numerical measures based on the intensity or proximity implied by the respondents. This was key to understanding how different relationship dynamics affect fertility intentions.

We standardized continuous variables such as frequency of face-to-face interactions and non-face-to-face interactions so that they contribute proportionally to the model’s learning. This normalisation helps the GNN to weigh the importance of different types of interactions without bias towards variables with larger ranges.

Besides making sure the GNN got good input, this thorough data preparation allowed it to capture the intricate patterns of social influence on fertility intentions within networks. I set up everything for the next stages of analysis by controlling every aspect of data management which made the results more valid and reliable in the end.

3.3 ETHICAL AND LEGAL CONSIDERATIONS OF THE DATA

The data in this study come from Stulp et al. (2023). All subjects gave informed consent and were fully informed of their rights, including the right to withdraw at any time. Personal data was anonymized and GDPR compliant and was approved by an ethics committee. I follow the ethics and data usage statements of the original study.

4. METHODS

4.1 TRANSLATION OF THE RESEARCH QUESTION TO A DATA SCIENCE QUESTION

The research question for this thesis has been narrowed down to focus on the analytical and technical capabilities of Graph Neural Networks (GNNs) for predicting reproductive intentions in women's social networks. The question is: "How can we optimise the architecture of Graph Neural Networks to predict fertility intentions better given the features and interactions in women's social networks?" This rephrased research question highlights the need for more advanced machine learning methods that can handle the nuances and complexity of social data.

Relationships are complex and individual choices are hard to predict. Improving how these models gather and analyse the complex network of social ties and individual attributes is part of optimising the GNN architecture. This means figuring out how to arrange the network graph's layers, node embeddings and connectivity patterns to reflect and forecast the impact of different social structures.

4.1.1 Preprocessing

To ensure methodological consistency and robustness the first step of the research involved thorough data cleaning according to the protocols of Stulp (2023). The dataset needed to be cleaned to be ready for graph-based modeling, addressing common issues like inconsistent or missing data and misformatted node properties or edges. Instead of using typical imputation methods the preprocessing strategy used established groups or categories to handle missing values and preserve the structure needed for network analysis. This prevented the distortions caused by scaling continuous variables like age and frequency of encounters, categorizing them into well defined groups that reflect the underlying social dynamics. This is supported by Zhou et al. (2022) and Goel et al. (2022) who highlight the importance of architectural optimizations and good data handling in Graph Neural Network (GNN) applications.

Table 1 shows the specific conditions and corresponding values used during data preparation. By categorizing and standardizing the main attributes this table is a reference for the methods used to clean the data for graph based analysis. These steps are necessary to align the dataset with the structural and analytical requirements of Graph Neural Networks and to keep the data robust and interpretable in the network context

Attribute	**Condition**	**Preprocessed Value**
Age	50+	50
	18-	18
	Numeric directly provided	As provided
	Missing or undefined	0
Number of Children	Expecting first child	0.5
	More than 5	5
	Numeric between 0 and 4	As provided
	Missing or “I don’t know”	0
Sex	Female	1
	Male	-1
Happiness Related to Children	“I don’t know” or child not yet born	0
	Happiness decreased or values 1, 2	1
	Happiness increased or values 4, 5	2
	Happiness remained the same or value 3	3
Child-Free Preference	Prefers to remain childless	0
	Wishes to have children	1
	Undecided or ambiguous	2
Friendship Status	Yes, is a friend	1
	No	-1
	Missing or unclear	0
Relationship Type	Textual description	Extract first numeric value
	Numeric directly provided	As provided
	Missing or unclear	0
Face-to-Face Interaction Frequency	Valid numeric entry	As provided
	Missing or unclear	0
Non-Face-to-Face Interaction Frequency	Valid numeric entry	As provided
	Missing or unclear	0
Helpfulness in Child-Rearing (help)	Valid numeric entry	As provided
	Missing or not applicable	0

Table 1: Details the specific conditions and transformations applied to key attributes to prepare them for GNN analysis.

4.1.2 Feature Selection

I focused on finding and using characteristics that alters and ego had in common during the feature selection stage of my research so I could get a detailed look at how individual characteristics and social connections impact fertility goals. This was based on the idea that shared characteristics, such as age, number of children and relationship status could be the mediating factors in how social networks affect personal fertility choices. I wanted to clarify the complex interactions between social factors and personal predispositions within the network by looking at these common traits. With a better understanding of the underlying social dynamics at play this allowed me to get a more targeted look at how similar or different characteristics of people within the same network could affect their opinions and choices around fertility.

- Age (age): Age is a key demographic variable for fertility decisions. It was included as a continuous variable to capture the changes in fertility attitudes and behaviors as women get older.

- Number of Children (num_children): This measures the number of children an individual has. It's a direct measure of past fertility behavior and future intentions. This variable was included as both a continuous count and categorical to test thresholds at which intentions change.

- Gender (sex): Although the focus is on women, understanding the gender dynamics within the social network can give insights into how gendered expectations influence fertility decisions. This categorical variable differentiates responses and interactions within the network.

- Child-Free Preference (child_free): This captures whether an individual wants to remain childless. It's a key aspect of modern fertility studies, reflecting personal life choices and societal shift in attitudes towards parenting. It was included as a binary variable.

- Friendship Status (friend): This indicates whether someone in the respondent's network is a friend. Friendship can have a big impact on personal decisions through emotional and informational support. It was operationalized as a binary variable to distinguish between friends versus acquaintances or family.

- Face-to-Face Interaction Frequency (f_to_f): This measures how often the respondent meets someone face-to-face. Regular physical meetings can strengthen social bonds and more significant influence. This was included as an ordinal variable.

- Relationship Type (relationship_type): Captures the type of relationship between the respondent and others in their network (e.g., family, coworker, friend). Different types of relationships can have different influences on fertility decisions. This variable was modeled as categorical.

- Perceived Helpfulness (help): Whether the respondent can ask someone in their network for help. This was included as a binary variable.

- Non-Face-to-Face Interaction Frequency (non_f2f): How often the respondent interacts with someone through digital means (e.g., phone calls, emails, social media). As digital communication becomes more common, its role in shaping social influence grows. This variable was also treated as an ordinal scale.

- Increased Happiness Post-Children (happiness): How happy the respondent feels after having children, a big factor in wanting more children.

4.1.3 Model Architecture

The social dynamics in the network are captured perfectly by the Graph Neural Network (GNN). It had 3 layers of graph convolutions, each of which could change a node’s representation by aggregating features from its immediate neighbours. This is how social influence spreads in the network. I added dropout layers to prevent overfitting and ReLU activation functions to allow the model to see patterns in the input. This ensures the model's results are valid and generalizable across different parts of the network.

Training The Model

Training was done carefully to get the best out of the model. I used cross-entropy loss function to classify multi-class outcomes like fertility intentions. Adam optimizer was chosen because it's good at handling sparse gradients and adjusting learning rates. Model was trained for a large number of epochs to get efficient learning and model fusion. Batch size and other params were tuned based on initial performance.

Evaluation Metrics

To evaluate the GNN comprehensively, I used multiple metrics. The class imbalance in the dataset made precision, recall and F1 score more important. These metrics gave a fair evaluation of the model's accuracy and ability to handle minority classes. A breakdown of the model's performance across different fertility intention categories was also done using a confusion matrix which highlighted areas of the model that needs improvement.

All my methodological choices were made with my thesis's main question in mind: to study how social networks affect fertility intentions and use graph-based ml to see the details in the data.

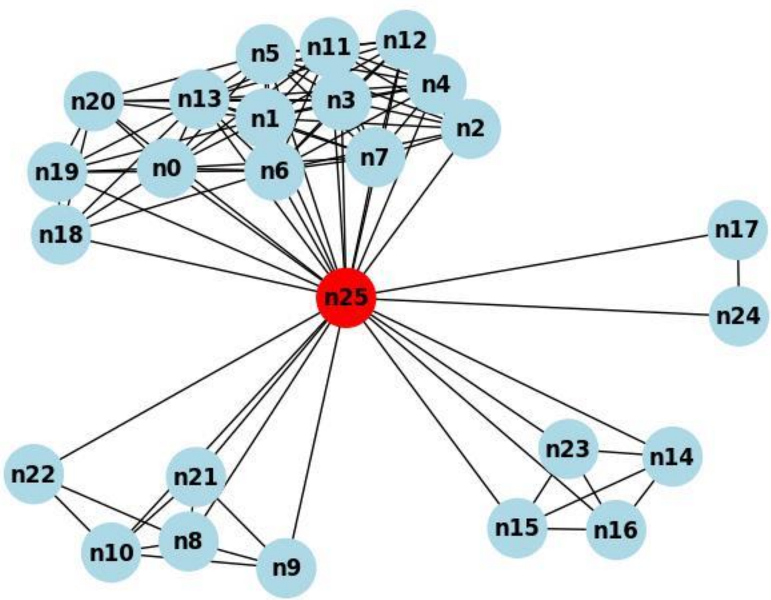

Figure 1: The figure depicts a network graph centered on a primary respondent (red node) and illustrates their connections within a broader social network.

4.2 MOTIVATED SELECTION OF METHODS FOR ANALYSIS

I chose my analytical methods for my research because I needed to document exactly and to look at the complex relationships within social networks that affect fertility goals. Because they can handle and analyse data that has relational structure like the networks of people and their social ties Graph Neural Networks (GNNs) were my choice for data management. This was especially fitting given the emphasis on the correlation between different fertility goals and shared attributes between alters and egos within these networks.

Because it can process data directly in graph format where nodes are individuals and edges are the relationships between them, Graph Neural Networks are perfect for this. GNNs are good at combining node level attributes with the overall network architecture so you can analyse everything at once, individual attributes and their context. This allows the model to capture complex structures of influence and interaction that are not visible with traditional analytical tools which is important to understand how social effects within a network can affect an individual’s decision on fertility.

GNNs were chosen because of their ability to learn from big and complex data and their flexibility in handling multiple layers of information. I can describe and analyse the social dynamics better by using a method that fits the complexity of social network data. Also feature learning can be done automatically with GNNs which is very helpful in social science research where relationships and influences are not always obvious.

Also the robustness and reliability of the findings was ensured by using cross-validation method. By doing this I was able to test the GNN model on several network subsets and be confident that the patterns I saw were not just an artifact of the sample or an overfitting model.

The analysis of individual attributes and social ties in determining fertility goals within networks was made possible by Graph Neural Networks (GNNs).

4.3 MOTIVATED SETTINGS FOR SELECTED METHODS

To get the model to show and analyze the social dynamics of fertile intentions I had to fine tune the settings. Because the Adam optimizer is great at handling sparse gradients which is common in graph data I chose it. I dynamically adjusted the learning rate for every parameter and watched the performance metrics during the early training phases to make iterative changes to the learning rate and other hyperparameters to get the model to converge.

I used regularization techniques like dropout to combat the issues caused by the model's depth and the complexity of the network data. During training the model is forced to learn features that are independent of any particular subset of the data by randomly dropping a section of the neurons. This helps the model to generalise across different network configurations and be more robust and adaptable.

Since cross-entropy loss function works well for multi-class classification tasks which is a task that matches the categorical nature of my study's outcome, fertility intentions it was an excellent choice. This probabilistic method ensures each category of female desire is well represented and forecasted so I can have a more nuanced evaluation of the model's performance across three classes ("Yes, I want a child", "No, I don't want a child", "I don't know")

A strong cross-validation was needed to evaluate the model's stability and performance. To make sure each fold acted as a test set at different times and the rest as training set, the data was divided into 5 folds. This also helps to evaluate the Graph Neural Network's performance across the network and prevent overfitting and get reliable and reproducible results.

The model was trained for enough epochs with early stopping in place to stop training if no improvement in validation performance was seen. This ensures the quality of learning. This approach balances the model training and depth of learning and avoid unnecessary computation and overfitting.

By making these choices the project aims to provide very accurate and actionable insights so I can better understand how social interactions in the network affect fertility intentions. This way every part of the model contributes to the complex relationship between individual characteristics and social factors in determining reproductive outcomes.

The source code and datasets for the analysis are available in my GitHub repository (Diamantidis Dimitrios, 2024).

5. RESULTS

In this study we used Graph Neural Networks (GNNs) to see how fertility intentions are affected by individual and network level variables in social networks. The analysis was split into three stages, each to improve the model and fix issues that arose from the first results. The main goal was to improve the GNN model to classify people according to their self reported fertility intentions which can be ambiguous ("I don't know"), not interested ("I don't want a child") or interested ("I want a child").

To set a baseline and detect any early biases or errors in the forecast we first deployed a basic GNN model. This step was important to set the ground for later optimisations and to see how the model behaves with imbalanced raw input. The first results were a problem, despite the model could interpret complex network data, it was favouring the class that was less represented in the sample, predicting "I don't know" much more than statistically expected. This was unexpected and showed that we need to intentionally change the way the model handles feature processing and data imbalances.

I went through each step methodically to fix these issues, using advanced machine learning techniques like over-sampling and cross-validation, fine tuning the model and the data. The goal was to keep the model to classify the majority classes while being sensitive to the minority class so to have a more accurate and balanced representation of the social effects on fertility intentions.

5.1 INITIAL BASIC GNN MODEL IMPLEMENTATION

We first established a baseline of how well the initial model could predict fertility intentions within social networks by the first phase of the Graph Neural Network (GNN) implementation. In this stage we built a simple GNN architecture of 3 graph convolutional layers. These layers were meant to process and learn from the complex relational data within the participant network.

We expected the model to be able to distinguish between the main fertility intentions despite its simplicity. But the results of the first model deployment showed us some issues. The model was predicting the "I don't know" category and it was biased. This was unexpected because the least common category in the dataset had an anomaly between the model's predictions and the actual data.

Some of the feature combinations tested had the following attributes: "number of children", "child free", "friend", "face to face", "relationship type", "help". This combination had an F1 score of 0.1835 which is very poor. The confusion matrix shows the problem clearly, the model can't predict any of the other classes "I want a child" or "I don't want a child". Especially this combination's confusion matrix shows that out of all the "I don't know" predictions, none of them were correct, which means the model completely misdiagnosed the other two most common categories (Figure 1).

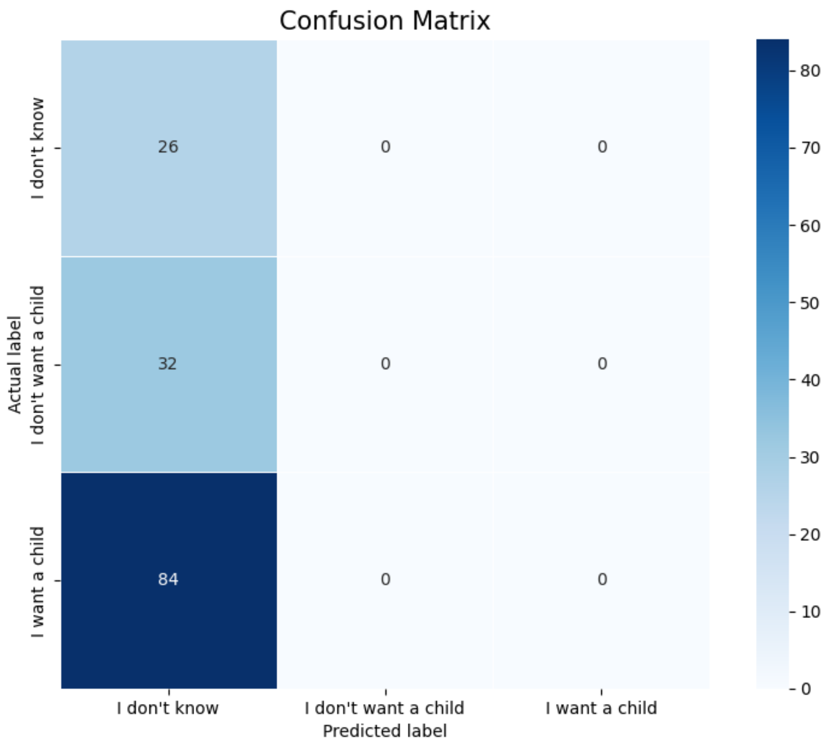

Figure 2: The confusion matrix shows the GNN model's predictions, incorrectly classifying all responses as "I don't know" for the three different categories.

This preliminary investigation showed that the model needs more work. The unbalanced training data or insufficient feature representation that could not capture the subtleties of people's decisions on fertility might be the reason for the biased predictions which means possible issues with how the model is learning from the data. Knowing these weaknesses was important to guide the next stages of the analysis which includes applying more advanced methods to address data imbalance and improve the model's ability to predict across the different scenarios in the dataset.

5.2 IMPROVED MODEL WITH OPTIMIZATION

A number of repeated changes were made to the model's parameters in order to improve its prediction performance during the Graph Neural Network (GNN) analysis's optimisation phase. In this phase, the model's ability to increase classification accuracy across all categories of fertility intention was assessed in light of the improved training strategies.

The model's performance significantly improved once optimisation techniques were used, however problems with class imbalances remained. The attribute combination number of children, child-free preference, friendship status, face-to-face interaction frequency, relationship type, perceived helpfulness produced the most illustrative result. With this arrangement, the F1 score increased to 0.2691 from lower starting points, indicating a little improvement in predicting accuracy. The accompanying confusion matrix, however, showed that the improvements were uneven across the classes. In particular, the matrix showed:

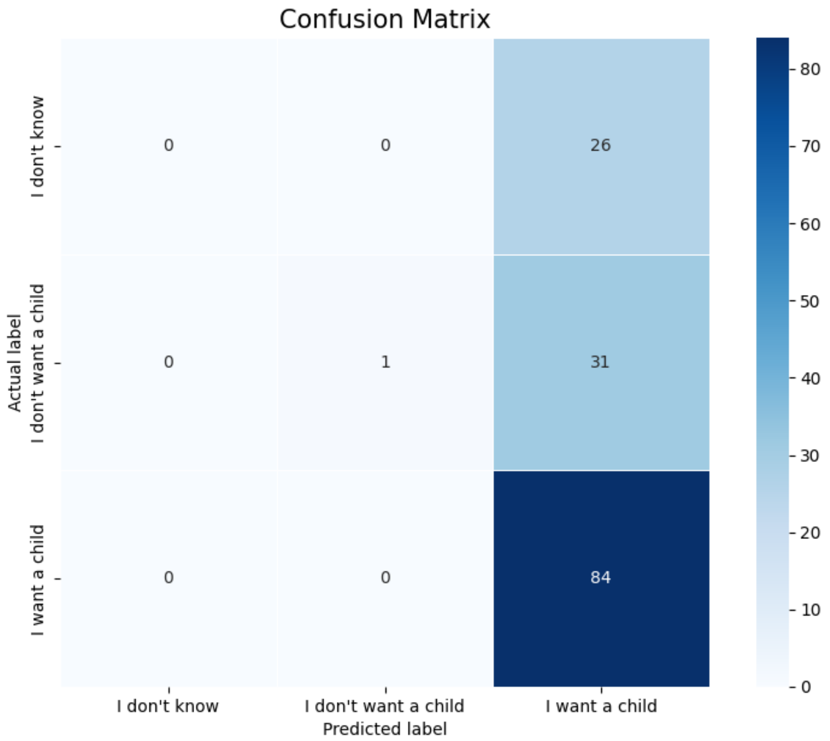

Figure 3: Confusion matrix of the GNN model with optimized parameters, using number of children, child-free preference, friendship status, face-to-face interaction frequency, relationship type and perceived helpfulness as attributes.

Other examined combinations also showed this pattern, though to differing degrees of precision. For example, a slightly lower F1 score of 0.2478 was obtained for a different collection of attributes age, number of children, gender, relationship type, perceived helpfulness, non-face-to-face interaction frequency, nonetheless, the confusion matrix still showed no right predictions for any of the classes across many test instances:

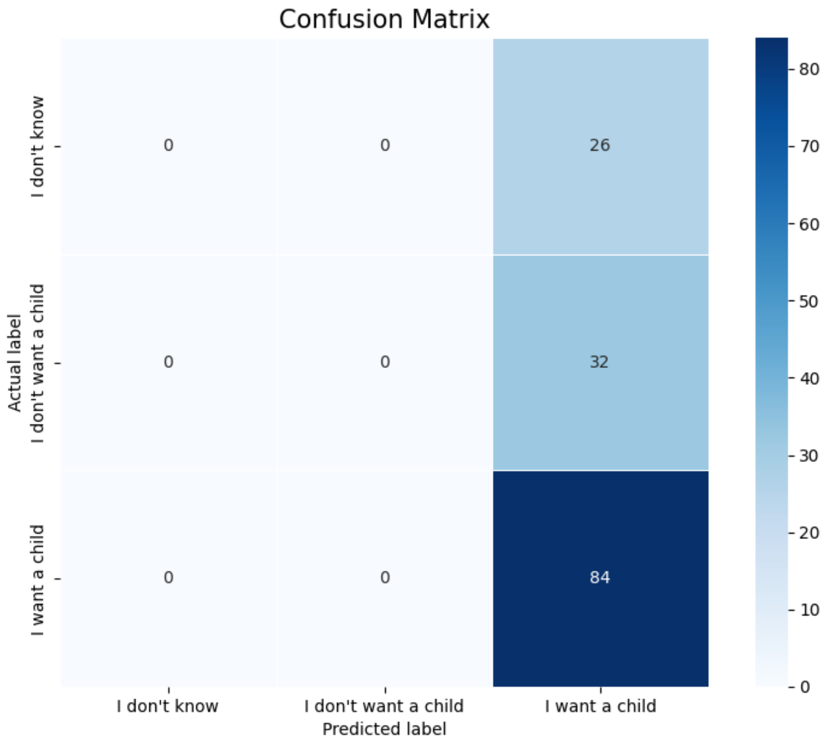

Figure 4: Confusion matrix of the GNN model with optimized parameters, using age, number of children, gender, relationship type, perceived helpfulness, non-face-to-face interaction frequency as attributes.

These findings highlighted an important point, even though the optimisation procedure helped to modify the model's weight in response to training failures, the model's focus readjustment was still insufficient to address the skew towards specific classes. The model's persistent random prediction pattern suggested that it still needed to be improved.

5.3 ADVANCED MODEL WITH OVER-SAMPLING AND CROSS-VALIDATION

In the final stage of the research, the Graph Neural Network (GNN) was trained to apply over-sampling and cross-validation to reflect and forecast the different fertility goals classes. This was to address the class imbalance that had been affecting the model’s previous estimates.

To balance the training dataset and reduce the bias towards the most frequent class, over-sampling was applied. The performance metrics showed significant improvements across different attribute combinations. The combination of number of children, child-free preference, friendship status, face-to-face interaction frequency, relationship type, and perceived helpfulness scored 0.4371 which is a big improvement from previous stages. The cross-validation results also supported the stability of this model configuration and showed consistent improvement in model reliability. Scores ranged from 0.4484 to 0.5004 across 5 folds.

This model iteration's confusion matrix showed a more evenly distributed prediction across all categories:

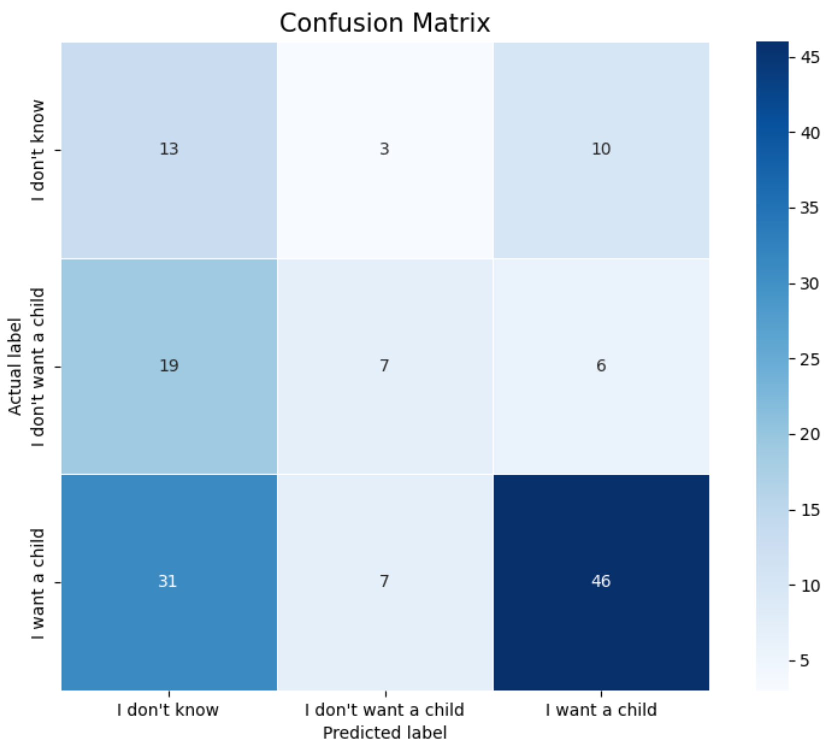

Figure 5: Confusion matrix of the GNN model with over-sampling and cross-validation, using number of children, child-free preference, friendship status, face-to-face interaction frequency, relationship type, and perceived helpfulness as attributes.

A score of 0.5033 was achieved with the additional attribute combination of age, number of children, gender, child-free preference, friendship status, relationship type, and non-face-to-face interaction frequency. The cross-validation scores were 0.4560, 0.5002, 0.4622, 0.4966, 0.4853 which show some variation but overall good performance. The confusion matrix also showed better distribution in every class:

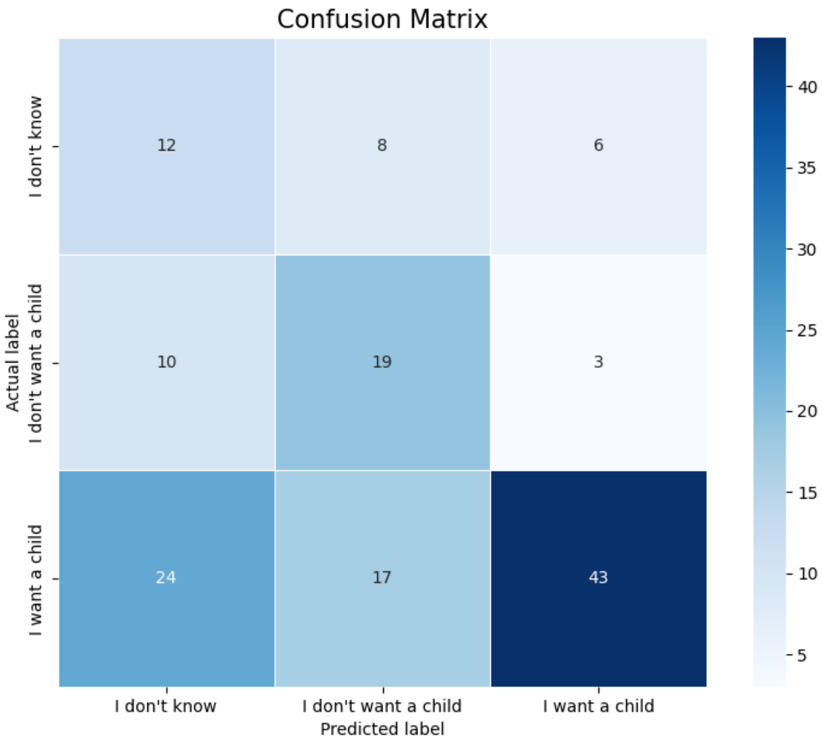

Figure 6: Confusion matrix of the GNN model with over-sampling and cross-validation, using age, number of children, gender, child-free preference, friendship status, relationship type, and non-face-to-face interaction frequency as attributes.

These results show a big improvement in the model's ability to handle the dataset's complexity. The ability to classify across all fertility intention categories showed how oversampling and cross-validation were integrated into the model training process. This stage not only corrected the biases found in previous models but also proved that when properly tuned and tested GNNs can capture subtle patterns in social network data.

So we built on that foundation and made some more structural changes to the GNN. We tried reducing the number of layers to see what was the optimal complexity for the dataset. Reducing to 2 layers was similar to the original 3 layer model, both around 50% accuracy. So reducing a layer doesn't hurt the network's ability to find patterns in the data. Increasing to 4-6 layers didn't improve performance, staying at 49-50% accuracy. So we hit a performance plateau in this layer range. Going to 8 layers dropped performance (f1-score) to 40% and the confusion matrix (Figure 7) showed misclassifications across all categories, likely overfitting.

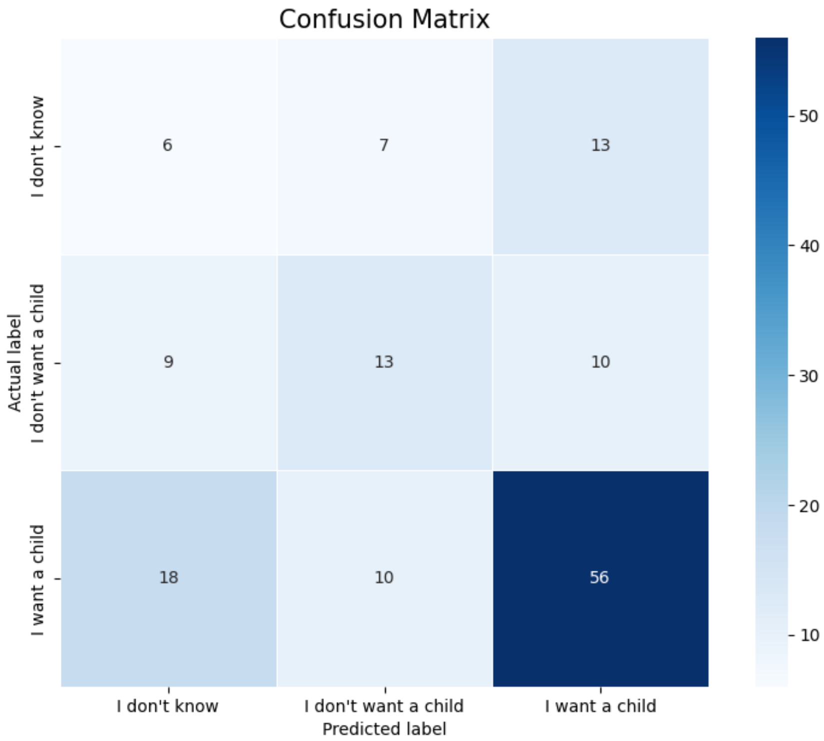

Figure 7: Confusion matrix for the GNN model with eight layers, using age, number of children, gender, child-free preference, friendship status, relationship type, and non-face-to-face interaction frequency as attributes.

So it looks like the GNN doesn't gain much from added depth beyond the initial 3 layers, so this might be the sweet spot for this task. This level of complexity is enough to process and represent the data without overfitting. This is the delicate balance in neural network design where we have to manage underfitting and overfitting to get the best performance.

Having achieved the highest F1-score to date with the final optimized model, the confusion matrix (Figure 6) gives us a good idea of how the model is performing on reproductive intentions within social networks. This matrix shows the model is very good at identifying people who definitely want to have children, 43 true positives in this category. This means expressions of wanting to have children are often clear and the model is picking them up well.

On the other hand the matrix shows challenges in the less clear categories. For those who are unsure about having children ("I don't know") the model only identified 12 true positives and misclassified 34 as wanting or not wanting children. This means the model is having trouble distinguishing between indecision and more clear reproductive preferences. Similarly for those who don’t want children the model got 19 right but misclassified 25 as belonging to this category, this means overfitting or misinterpretation of the signals people use to express not wanting children.

In summary the matrix shows the model is good at explicit intentions and bad at the more subtle aspects of human reproductive intentions. This means we need to improve the model to interpret and classify the ambiguous or subtle intentions in the social data.

6. CONCLUSION AND DISCUSSION

6.1 ANSWERING THE DATA SCIENCE AND RESEARCH QUESTIONS

This study tested whether GNNs can predict reproductive intentions in social networks. The results show that while GNNs can capture complex network relationships that influence individual decisions, they peak at 50% accuracy. This is in line with the challenges GNNs face when applied to social science data which is complex and skewed. As Goel et al. (2022) noted, training GNNs on complex datasets that don’t fit typical modeling frameworks. And as Bronstein et al. (2020) said, we need to integrate complex relational data into GNNs to make them more predictive.

Stulp et al. (2023) found that individual characteristics dominate over network variables in predicting fertility preferences using a data-driven LASSO regression approach. We differ in our approach but agree that while network effects are important, the core individual drivers play a bigger role in determining fertility outcomes than network-only models account for. Despite over-sampling and cross-validation to address class imbalance and improve predictions, our GNN models performed modestly. The advanced model stages showed some improvement in fairness and balance across categories but not in high F1 scores, so there’s a gap between GNNs and social network data, especially in predicting human behaviors like fertility intentions.

So GNNs give us some answers to reproductive intentions in social networks but there’s still much to be developed and fine-tuned. They need to handle the nuances and diversity of social data better. Future research should focus on model architectures, data preprocessing and feature sets that can capture deeper social influence. In summary, this study shows that GNNs can be used in social science but need significant innovation and adaptation to social data. The journey to high accuracy is long, so we need a rigorous method and continuous model evaluation.

6.2 DESCRIBING IMPLICATIONS FOR THE PROPER DOMAIN SETTING

Using Graph Neural Networks (GNNs) to predict fertility intentions in social networks may not have given us great results but it's a nice example of how we can apply advanced computational methods in social science. This work shows that GNNs can give us a deeper understanding of how social context affects individual decisions like fertility goals. But since we didn't get high accuracy, these insights should be considered as preliminary. Social scientists and policymakers should use these results when designing interventions informed by network dynamics but also be aware of the limitations of the current model to predict individual outcomes. For example, while family planning programs could be adapted based on network dynamics as suggested by this research, we should proceed with caution and make sure strategies are backed up by multiple sources of data to increase reliability and effectiveness.

6.3 DISCUSS ETHICAL IMPLICATIONS AND CONSIDERATIONS

There are also ethical issues raised by the application of powerful machine learning algorithms like GNNs especially those related to data handling and privacy. Strict steps were taken in this study to get informed consent and anonymize data, protect participant identities while allowing us to examine their social contacts thoroughly. But if more complex network data or integrated datasets are mishandled, re-identification is possible. So any future research using these kinds of techniques should strictly follow data protection ethical standards and be transparent about data use and storage policies. Also, we should be careful when interpreting machine learning results with regards to social behaviors to avoid making unwarranted generalizations or discriminating certain groups based on model results only.

In summary, this work shows the analytical capability of GNNs in social science and also reminds us to be cautious with ethics and domain specific considerations when we bring in cutting edge computational methods in human centered research. By weighing the pros and cons of these technologies, we can use machine learning to understand complex social phenomena while being ethical.

7. REFERENCES

1. **Stulp, G., Top, L., Xu, X., & Sivak, E. (2023).** A data-driven approach shows that individuals' characteristics are more important than their networks in predicting fertility preferences. Royal Society Open Science, 10(12), 230988.
2. **Oloulade, B. M., Gao, J., Chen, J., Lyu, T., & Al-Sabri, R. (2021).** Graph neural architecture search: A survey. Tsinghua Science and Technology, 27(4), 692-708.
3. **Wei, L., Zhao, H., He, Z., & Yao, Q. (2023).** Neural architecture search for GNN-based graph classification. ACM Transactions on Information Systems, 42(1), 1-29.
4. **Zhou, K., Huang, X., Song, Q., Chen, R., & Hu, X. (2022).** Auto-gnn: Neural architecture search of graph neural networks. Frontiers in big Data, 5, 1029307.
5. **Shao, Y., Li, H., Gu, X., Yin, H., Li, Y., Miao, X., ... & Chen, L. (2024).** Distributed graph neural network training: A survey. ACM Computing Surveys, 56(8), 1-39.
6. **Vesselinova, N., Steinert, R., Perez-Ramirez, D. F., & Boman, M. (2020).** Learning combinatorial optimization on graphs: A survey with applications to networking. IEEE Access, 8, 120388-120416.
7. **Peng, Y., Choi, B., & Xu, J. (2021).** Graph learning for combinatorial optimization: a survey of state-of-the-art. Data Science and Engineering, 6(2), 119-141.
8. **Liang, F., Qian, C., Yu, W., Griffith, D., & Golmie, N. (2022).** Survey of graph neural networks and applications. Wireless Communications and Mobile Computing, 2022(1), 9261537.
9. **Shi, S., Qiao, K., Yang, S., Wang, L., Chen, J., & Yan, B. (2021).** Boosting-GNN: boosting algorithm for graph networks on imbalanced node classification. Frontiers in neurorobotics, 15, 775688.
10. **Xia, R., Zhang, C., Zhang, Y., Liu, X., & Yang, B. (2024).** A novel graph oversampling framework for node classification in class-imbalanced graphs. Science China Information Sciences, 67(6), 1-16.
11. **Marmol Asenjo, A. (2023). Optimization and Cross-Validation of Graph Neural Networks for the Diagnosis of Alzheimer's Disease.**
12. **Goel, S., Shetty, A., & Goel, A. GNN: A Survey on Architectures and Optimizations.**
13. **Diamantidis, D.(2024). Optimizing Graph Neural Networks for Predicting Fertility Intentions in Social Networks [Pycharm]. GitHub.**

https://github.com/dimitrisdiam/Optimizing-Graph-Neural-Networks-for-Predicting-Fertility-Intentions-in-Social-Networks