MSc. Applied Data Science Thesis

Utrecht University

PREDICTING POSITIVE AND NEGATIVE TIES IN SIGNED NETWORKS

A Comparative Study on Online and Real World Networks

Bo Staals, 1037803

06 July 2023

UU Supervisor

Dr. Javier Garcia Bernardo

Second Supervisor

Dr. Eva Jaspers

Abstract

Network analysis is a growing area of research in various fields. While most existing studies focus on unsigned networks, this research explores the coexistence of positive and negative ties in signed online and real world networks. By combining topological features and social theories, this research investigates the gap in previous research. This leads to the following research question: *"To what extent can machine learning models predict positive and negative ties in online and real world networks?"*

The network analysis is carried out on two signed networks. The Wikipedia network represents the online network, where every user is allowed to vote for other users that request for adminship. The school network represents the real world network and is based on data collected in the "Children of Immigrants Longitudinal Survey in Four European Countries" project. The final datasets used for the classification model contain 4627 nodes and 122767 ties for the Wikipedia network and 4284 nodes and 27333 ties for the school network.

Snowball sampling has been performed to obtain subsets of the networks due to computational bottlenecks. The final feature set includes every connection between nodes, which can be either negative, not observed or positive. It also includes important metrics such as node-level, network-level, similarity-based, social theory and path-based metrics. After model selection, hyperparameter tuning and feature selection on the subsets, the final datasets are evaluated on the tuned Light Gradient Boosting Machine model.

The results show that the school network outperforms the Wikipedia network in terms of F1, recall and precision in all classes except the not observed class. Reasons for this difference could be the imbalanced datasets and different network structures. The Wikipedia network has biased predictions towards the not observed class while the school network shows consistent performance.

In conclusion, the real world network has a higher performance compared to the online network. We must keep in mind that comparing these networks is challenging due to the different network structure. Therefore, future research could compare more similar networks to obtain more generalisable results. In addition, it is important to overcome the problem of imbalanced data in order to obtain reliable and consistent results.

Keywords: Network analysis, signed networks, positive and negative ties, machine learning models, online and real world networks, Python.

Acknowledgements

I would like to express my gratitude to all those who have contributed to the successful completion of this research. First and foremost, I would like to thank my supervisor dr. Javier Garcia-Bernardo for his guidance and astute insights. He has provided invaluable feedback on my research and filled in the gaps in my computer science knowledge. I would also like to thank my second supervisor dr. Eva Jaspers for her interest in my research and for providing valuable data. I would also like to thank Elena Candellone for her valuable feedback sessions that have enriched this research. Finally, a thank you to Filip Chrzuszcz and Sofoklis Repopoulos for the great collaboration and valuable discussions.

Abbreviations

A	Adjacency Matrix
AA	Adamic-Adar
AUC	Area Under Curve
CILS4EU	Children of Immigrants Longitudinal Survey in Four European Countries
CN	Common Neighbours
CV	Cross Validation
LGBM	Light Gradient Boosting Machine
ML	Machine Learning
MLR	Multinomial Logistic Regression
OvR	One versus Rest
PR	Precision-Recall
RA	Resource Allocation
RfA	Request for Adminship
RFECV	Recursive Feature Elimination Cross Validation
ROC	Receiver Operating Characteristic
RW	Random Walk
RWWR	Random Walk With Restart
SKCV	Stratified K-Fold Cross-Validation
SNAP	Stanford Network Analysis Platform

Table of Contents

1 Introduction

1.1 Motivation and context

Network analysis is an emerging topic in many fields, including social science (e.g., analysing teacher-teacher relationships) [1], political science (e.g., analysing polarisation) [2], and health science (e.g., analyse spreading of MERS coronavirus) [3]. It helps answer questions about network structure, the formation of network ties, and the prediction of these network ties [4]. Most existing network analysis has been performed on unsigned networks. However, real world networks can be signed networks (see Figure 1) [5]. In signed networks, positive and negative ties coexist. The positive ties are formed by friendships, support and alliances between individuals. The negative ties are formed by conflicts, disagreements and bullying. Most research has only looked at the positive ties, as negative ties are often not observed [6]. However, there is a need for methods that examine the negative ties in addition to the positive ties [5]. This will provide a deeper real-world understanding of social dynamics in networks. Applications of signed network prediction include detecting criminal activity in terrorist networks [7], detecting fraudulent users in mobile phone networks [8], and identifying who is being bullied in school networks [9]. The, the aim of this research is therefore to predict positive and negative ties in social networks.

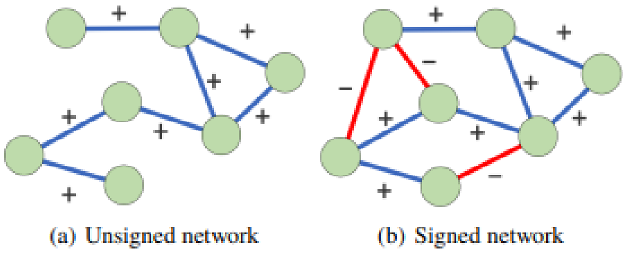


Figure 1: Unsigned networks (a) and signed networks (b) [5].

1.2 Literature overview

This literature overview explores the topics of networks, link prediction heuristics, and social balance and status theories.

1.2.1 Node-level and network-level metrics

Networks consist of nodes, which are objects or individuals, and links, ties or edges, which represent the relationship between nodes. For the sake of clarity, the term 'ties' is used in this research. Node-level measures identify the importance or centrality of a node in the network. Three important metrics are the *degree*, which assigns an importance score based on the number of ties, the *eigenvector*, which also considers how well connected a node is and the *PageRank*, which considers the direction of the ties in addition to the eigenvector

[10]. The degree can be divided into the *indegree*, which is the frequency of incoming nodes, and the *outdegree*, which is the frequency of outgoing nodes. This is often used to analyse directed networks. Network-level measures provide an overview of the characteristics of the network structure. The *network size* is the number of nodes in the network. In the *degree distribution,* the frequency of the degree is plotted to provide information about the connectivity of the network. The *density* indicates how connected the network is in terms of ties. The c*lustering coefficient* measures the extent to which your connected nodes are also connected to each other (e.g., your friends are also friends) [11].

1.2.2 Similarity-based metrics

Link prediction is the problem of predicting the existence of a tie between two nodes [12]. Besides node-level and network-level metrics, similarity-based algorithms are informative for link prediction. This is based on the principle of homophily in sociology; similar nodes are more likely to have a positive tie [13]. However, there is no evidence that this also applies to negative ties [5]. The algorithms follow the principle that each pair of nodes is assigned a similarity score s_{xy}. The not observed ties are ranked based on the s_{xy} score. Ties between similar nodes have a higher probability of existence [14]. The local similarity indices consider the neighbourhood of nodes and are computationally efficient. The global similarity indices consider the network structure but are computationally expensive [15]. The computational complexity varies from $O(2n)$ to $O(2n^2)$ [16]. Derr TS [5] shows that signed Random Walk with Restart (RWWR) gives the best performance (Area Under Curve (AUC) = 0.765) for sign prediction in the directed online Bitcoin-Alpha network. Research by Liben-Nowell D & Kleinberg J [12] shows that among similarity-based metrics, Adamic-Adar (AA) and Common Neighbours (CN) perform best on five physics networks. Furthermore, Feng et al. [17] investigate which topology features are important in terms of clustering. They conclude that for a low clustering real-world network, Superposed Random Walk (RW) is the best choice (precision = 0.03, AUC = 0.67). Otherwise, for a highly clustered real-world network, Resource Allocation (RA) has the best performance (precision = 0.57, AUC = 0.96). Finally, another paper by Zhou T et al. [18] shows that RA has the highest performance over six networks from different domains.

1.2.3 Social theory metrics

The two most significant social theories in signed networks balance and status. The social balance theory was introduced by the social psychologist Fritz Heider in 1946 [19]. The social status theory was introduced by the sociologist Weber and later developed by the computer scientist Leskovec. [20] in 2010.

The *social balance theory* states that "the friend of my friend is my friend" and "the enemy of my enemy is my friend" (Figure 2) [19]. A signed network is balanced if all triads contain an even number of negative ties. However, signed networks in the real world are rarely completely balanced [5]. The ratio of balanced and unbalanced triads is used to calculate the degree of balance [21, 22]. Leskovec et al. [20] state that balance theory is only applicable to undirected signed networks.

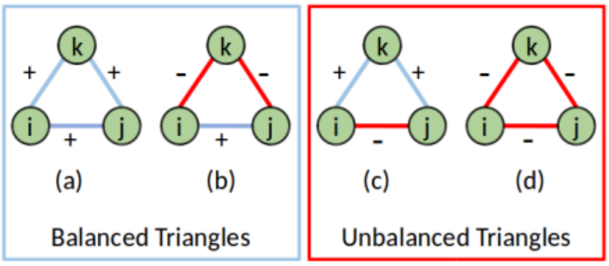


Figure 2: The social balance theory for the signed network triad [5].

Social status theory takes into account popularity and status (Figure 3). Nodes have positive ties to nodes with higher status and negative ties to nodes with lower status. Unlike social balance theory, status theory is applicable to directed signed networks [20]. Research from Derr TS [5] state that the social status theory can be implemented by the eigenvector centrality and the weight of positive and negative ties, see formula:

$$status = \sum (eigenvector * +1) - (eigenvector * -1).$$

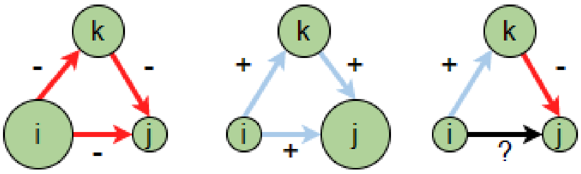


Figure 3: The social status theory for the directed signed network triad [23].

The social balance theory and the social status theory give contradictory results in directed signed networks. Consider the following situation where A has a positive tie to B and B has a positive tie to C. According to the social balance theory, C has a positive tie to A, but the social status theory assumes that this is a negative tie [20]. This is because balance theory considers undirected networks and status theory considers directed networks. Research by Leskovec et al. [20] suggests that status theory predictions are better than balance theory predictions in both online Epinions and Wikipedia networks. Robert West et al. [23] state that a mode incorporating both balance and status theory can reach AUC = 0.82 for the Wikipedia network. Implementing a text-based sentiment model in addition to the theories can increase the AUC up to 0.89. Another article [5] discusses that using the status and balance theories in addition to centrality measures leads to better prediction results.

1.2.4 Path-based metrics

Path-based metrics, which quantify the similarity between nodes based on the shortest paths connecting them, can also provide important information for link prediction [16]. Derr TS [5] discusses that the signed Katz achieves a high performance (AUC = 0.69) for the online Slashdot network. Kai-Yang Chiang et al. [24] suggest that the limitations of social balance theory in directed networks can be reduced by adding higher order k-cycles (e.g., 2-cycles have 4 configurations and 3-cycles have 16 configurations). The results show that up to k = 5 the accuracy increases and the false positive rate decreases. However, Wang P. [16] shows that considering longer cycles/paths is only useful if there are not many short paths. Furthermore, Liben-Nowell D & Kleinberg J [12] show that the Katz feature predicts the best (16% correct) on five physics networks, suggesting that there is still much room for improvement.

1.2.5 Link prediction methods

Three types of link prediction methods can be distinguished. First, *feature-based classification* is a supervised classification problem. Each pair of nodes is labelled positive if there is a tie and negative otherwise. The metrics discussed earlier can be used as features for the model, in addition to the non-topological features. The supervised models can improve the precision, but it is computational expensive [25]. Second, the *probabilistic graph model* assigns a probability value to each pair of nodes. This model achieves better accuracy performance compared to the classification model [26]. Finally, *latent feature methods* such as matrix factorisation can be used to predict ties. The basic idea is to factorise the adjacency matrix (*A)* into a low-rank latent-embedding matrix *Z* and its transpose. As the latent features in the network are captured, hidden connections can be identified [27]. This method can be combined with the other methods and optimises the AUC [28].

1.3 Research question

This research aims to address a gap in previous studies by investigating signed networks where positive and negative ties coexist. By combining topological features, social theories, and comparing online and real-world networks, this research will predict positive and negative ties using feature-based machine learning (ML) models. The research question is:

To what extent can machine learning models predict positive and negative ties in online and real-world networks?

2 Data

2.1 Data preparation for analysis

Data analysis, as discussed in the next chapter, is performed on two signed networks. See Figure 4 for a schematic overview of the networks.

The *Wikipedia network* represents the online network and is built around the Request for Adminship (RfA). From 2003 to 2013, data was collected on community members voting for a Wikipedia editor to become an administrator. The network consists of positive, negative and neutral ties, where neutral ties are treated as not observed ties [29]. The network was previously analysed by Leskovec et al. in 2010 [6] and 2014 [23]. In this research, a total of 12648 (6.4%) neutral ties are excluded from the analysis, as the aim is to predict positive and negative ties. A total of 7531 (4.1%) ties were duplicates, including 2882 ties with a conflicting sign. This means that both A → B is negative and A → B is positive. For these duplicates, the most recent vote is kept. Furthermore, the network consists of 12988 (7.3%) reciprocal ties, which means that both A → B and B → A exist. Therefore, the network is treated as undirected in order to extract the path-based and similarity-based features. A similar approach was used in the research of Leskovec J [20].

The *school network* represents the real world network and consists of data collected in the Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU) project. The main aim of this project is to investigate the internal processes leading to intergenerational integration. Data from four European countries have been collected in three waves, with wave 1 taking place at the age of 14 [30, 31]. For this study, the focus is on data from wave 1 (school year 2010/2011) from the Netherlands. This dataset consists of 100 schools, 222 classrooms and 4363 students [30]. The "Youth Classmate Questionnaire" is used to extract features that indicate a positive or negative tie between classmates. The following features are selected: "best friend" (+), "not want to sit by" (-), "often spend time with outside school" (+), reversed "sometimes mean to you" (-), "sometimes do homework with" (+) and "sometimes mean to" (-).[1] A total of 214 publications are based on CILS4EU data. Missing values are due to responses that could not be coded (other missing), no response due to a filter question (not applicable) and no response given (no answer). After transforming the dataset into a tie list dataset, the dataset only contains missing values in the 'classid' and 'schoolid' columns, as 3503 students are selected by other students but not interviewed themselves. A total of 13849 (31.6%) ties are duplicates, including 3612 ties with conflicting signs. For these duplicates, the vote from the most important features (best friend for positive and not want to sit by

[1] Positive ties are indicated with (+) and negative ties are indicated with (-).

for negative) is kept. Furthermore, the network consists of 16480 (55.1%) reciprocal ties. For the same reason as above, the network is considered undirected. Although students should only select students from their own class, 2583 (8.6%) students select students from other classes. These ties are deleted.

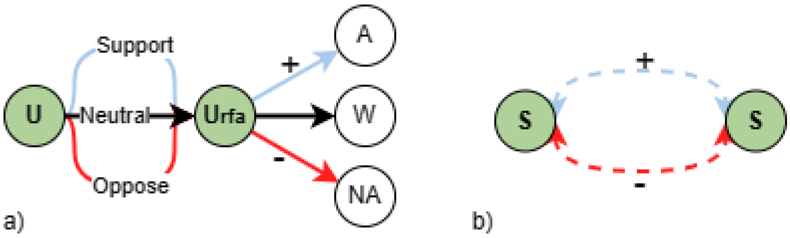


Figure 4: a) Wikipedia RfA network. Any user (U) can vote supporting, neutral or opposing for an user that requested for adminship (Urfa). The possible outcomes are either successful adminship (A), not successful adminship (NA) or withdrawal (W). b) The school network, where every student filled in a questionnaire. The questions ´best friend´, ´often spend time with outside school´ and ´sometimes do homework with´ are treated as positive ties. The questions ´not want to sit by´, ´sometimes mean to you´ and ´sometimes mean to´ are treated as negative ties. Ties can be either reciprocated, one-way or not existing.

2.2 Selected data exploration results

After data preparation, the Wikipedia network contains 11259 nodes and 178096 ties and the school network contains 4284 nodes and 27333 ties (see Table 1). Due to computational bottlenecks, the analysis for the Wikipedia network is performed on a subset. The subset consists of 4627 nodes and 122767 ties. A subset of the networks has been visualised in Appendix III to provide insight to the network structure.

Table 1: Exploratory data analysis for the Wikipedia network and subset and school network.

	Wikipedia	Wikipedia (subset)	School
Nodes	11259	4627	4284
Ties	178096	122767	27333
Positive ties	139510	97650	15636
Negative ties	38586	25117	11697
Reciprocal ties	12988	2686	16158
+ | +	10939	1692	10886
+ | -	834	432	822
- | +	834	432	822
- | -	381	130	3628

2.3 Ethical and legal considerations

The Wikipedia network is an open source network, extracted from the Stanford Network Analysis Platform (SNAP)[2]. The school dataset is not open source, but there is an option to request access[3]. Participants are anonymised to protect privacy and data is not shared.

[2] SNAP: Network datasets: Wikipedia Requests for Adminship (with text) (stanford.edu)
[3] Data (cils4.eu)

3 Methods

3.1 Translation of research question to data science question

To predict the positive and negative ties of the Wikipedia and school networks, the data is fed into a feature-based classification model. The two networks consist of the same information: the source of the vote, the target of the vote and the vote sign. This information includes essential node-level and network-level metrics that, together with similarity-based metrics and social theory metrics, are crucial for predicting ties. In addition to the existing positive (+1) and negative (-1) ties, the not observed ties (0) are added using the A. Due to computational bottlenecks, different classification models are first tested on a subset of the network to select the best classification model, hyperparameters and feature set. The best settings are then used in a final model.

3.2 Motivated selection of methods and settings

Snowball Sampling

After preparing the networks for analysis, a subset is created using snowball sampling. This sampling method is chosen because it preserves the network structure and is cost effective [32]. Previous research has also shown that snowball sampling is less biased than traditional random sampling [33]. Ten seed nodes and two layers are selected for the Wikipedia subset and ten seed nodes and five layers are selected for the school network. The motivation for the different settings is that in the Wikipedia network everyone votes for a user and in the school network everyone is restricted to voting for students within their class. For the final Wikipedia subset, 2000 seed nodes with two layers are selected to increase the sample size.

Feature set

The feature set baseline consists of the flattened A, which takes into account each connection between two nodes and its voting sign; 1 = positive, -1 = negative and 0 = not observed. To improve model performance, path-based metrics up to k = 5 are included, based on the results of Kai-Yang Chiang et al [24]. These features are derived from the undirected network to account for reciprocal ties. Before obtaining each higher order path, the diagonal of the previous A^k is removed to avoid bias from the degree feature [24]. Node-level and network-level metrics are then incorporated into the feature set. Each node-level metric includes two features, one for the 'source' node and one for the 'target' node. This can problems of multicollinearity, leading to bias in feature importance and overfitting. In addition, important similarity-based metrics are incorporated, requiring the network to be undirected. The selection of metric features is based on relevant literature (1.2) and availability within the networkX package. Finally, the feature set includes the social theory metrics. The path-based metric A² captures next to the nearest neighbour information and the balance between positive and negative ties, thus representing the social balance theory. The social status theory is implemented using the formula proposed by Derr T.S. and using node-level metrics [5]. A full description of the feature set can be found in Appendix I.

Model Selection

Stratified K-fold cross-validation (SKCV) is used to select the best performing machine ML model. SKCV uses stratified sampling in k folds to ensure the class frequencies. Specifically for the school network, the folds are created based on the 'classid'. S. Prutsky et al. [34, 35] discuss that the method is reliable and gives a more accurate estimate of performance in a health study. In addition, SKCV prevents some ties from being selected more often than others compared to random sampling [14]. In addition, class weights are implemented as the classes are highly imbalanced. This paper compares two models using SKCV, using the best performing model (highest F1 macro) for the analysis on the final datasets. The first model, Multinomial Logistic Regression (MLR), is a supervised learning model that is often used for classification problems. The second model, Light Gradient Boosting Machine (LGBM), is based on a tree-based algorithm that uses a gradient-based approach to optimise the model performance. It is known to be computationally efficient on large datasets [36]. These models are compared to see if additional complexity improves the model performance.

Hyperparameter Tuning

Then the hyperparameters are tuned. Weerts H. [37] discusses that several studies show that the tuning of hyperparameters is important to achieve a higher performance of a model. Therefore the hyperparameters are tuned using GridSearchCV. The combination of cross validation (CV) and grid search leads to more meaningful results [38].

Feature Selection

Next, Recursive Feature Elimination Cross Validation (RFECV) is used to select the important features. RFE reduces bias in the results by iteratively deleting insignificant features until the desired results are obtained [39]. In combination with CV, it can improve the performance of the model [40]. Feature selection is performed using the F1 macro scoring metric.

Model Performance Metrics

The final model obtained is then used to predict positive and negative ties on the final networks using SKCV and class weights to account for the imbalanced ties. The model performance is evaluated using the precision, recall, F1 and AUC, as they account for imbalance in networks (except AUC) [14]. An overview of the features is in Appendix II.

4 Results and Discussion

Model Selection Results

A robust model is created by combining SKCV, GridSearchCV and RFECV on the network subsets. Both the Wikipedia and school subsets show that LGBM (0.69 and 0.85 respectively) has a higher macro F1 compared to the MLR model (0.39 and 0.48 respectively). This suggests that the LGBM model has a better performance in terms of correctly identifying instances of all classes. Therefore, the LGBM model is selected for the final model analysis. After evaluating these results, the hyperparameters of the LGBM model are tuned and the feature set is determined. An overview of the selected hyperparameters and features is given in Tables 2 and 3.

Table 2: Selected hyperparameters for the LGBM model, determined using GridSearchCV (k = 5).

LGBM model	Wikipedia	School
n_estimators	200	200
learning_rate	0.01	0.01
max_depth	7	7
num_leaves	31	15

Table 3: The selected feature set for the Wikipedia network and school network, determined using RFECV (k = 5, tuned LGBM model, scoring = macro F1). For a more in-depth feature explanation, see Appendix I.

	Wikipedia	School
A2 (paths of length 2)		X
A3 (paths of length 3)	X	X
A4 (paths of length 4)		X
A5 (paths of length 5)	X	X
OUTDEGREE SOURCE		X
NEGATIVE INDEGREE SOURCE		X
EIGENVECTOR SOURCE		X
CLUSTERING SOURCE		X
NEGATIVE INDEGREE TARGET		X
CLUSTERING TARGET		X
SALTON		X
HUB PROMOTED INDEX		X
SOCIALSTATUS TARGET		X

Data Preparation Results

As mentioned before, the feature set baseline considers all connections between nodes minus the connections between the same node (e.g., node 1 → node 1). The final dataset of the Wikipedia network consists of 21402442 ties (0.2% negative, 98.9% not observed and 0.9% positive). In the school network, only the possible connections within each classroom are considered. The final dataset of the school network consists of 87850 ties (21.4% negative, 56.2% not observed and 22.4% positive ties). The results of the raw data analysis are shown in Appendix IV.

Model Performance Metrics

The model performance across all classes for both networks is shown in Figure 5. Figure 5a shows that for the *Wikipedia network,* the F1 score is the highest for the not observed class (0.91), indicating accurate classifications for this class. However, the model performs poorly for the negative (0.03) and positive (0.26) classes. The precision score follows the same pattern, indicating that the instance classified as 'not observed' is always correct (1.00). However, there are many false positives for both the negative and positive classes. The recall score is high for all classes, indicating that the model correctly identifies positive instances for each class. Figure 5b shows that for the *school network* the F1 score is high for all classes (0.75 – 0.80), indicating good performance across all classes. The precision score is highest for the not observed class (0.88), but moderate for the negative (0.67) and positive (0.72) classes. Recall is highest for the negative (0.84) and positive (0.89) classes. The not observed class has a slightly lower recall of 0.72. This indicates that there are few false positives and false negatives across all classes. Comparing this research with the research from Feng et al., [17] we see that the precision for the real world school network (0.67 – 0.88) is higher than the precision they found in their real world network (0.57). It is difficult to compare this research because different networks are used.

Comparing the two networks, the school network outperforms the Wikipedia network in terms of F1, recall and precision for all three classes except the not observed class. This implies that the model gives more accurate classifications for the school network. A possible reason for these results could be the imbalanced datasets due to the heterogeneous networks. In the Wikipedia network, the model is biased towards predicting the not observed class. However, this class imbalance problem is what we would like to overcome, and it failed for the Wikipedia network [16]. Another reason may be that the network structure is different for the two networks. The Wikipedia network has 10 years of data and the school network has 1 year, but both networks are treated statically rather than dynamically. In addition, in the Wikipedia network everybody can be connected while in the school network only students within each classroom can be connected. The voting process is also different; non-anonymous for the Wikipedia network, which can lead to herding bias [41], and anonymous for the school network. A final likely reason is that the networks are influenced by the method of feature selection. Two features are selected for the Wikipedia network and ten for the school network, so the predictive power may be higher for the school network. On the other hand, including too many features could also lead to overfitting in the school network. Also, the Wikipedia network might have had a lot of noise, which made the feature selection technique less effective, resulting in a lower performance.

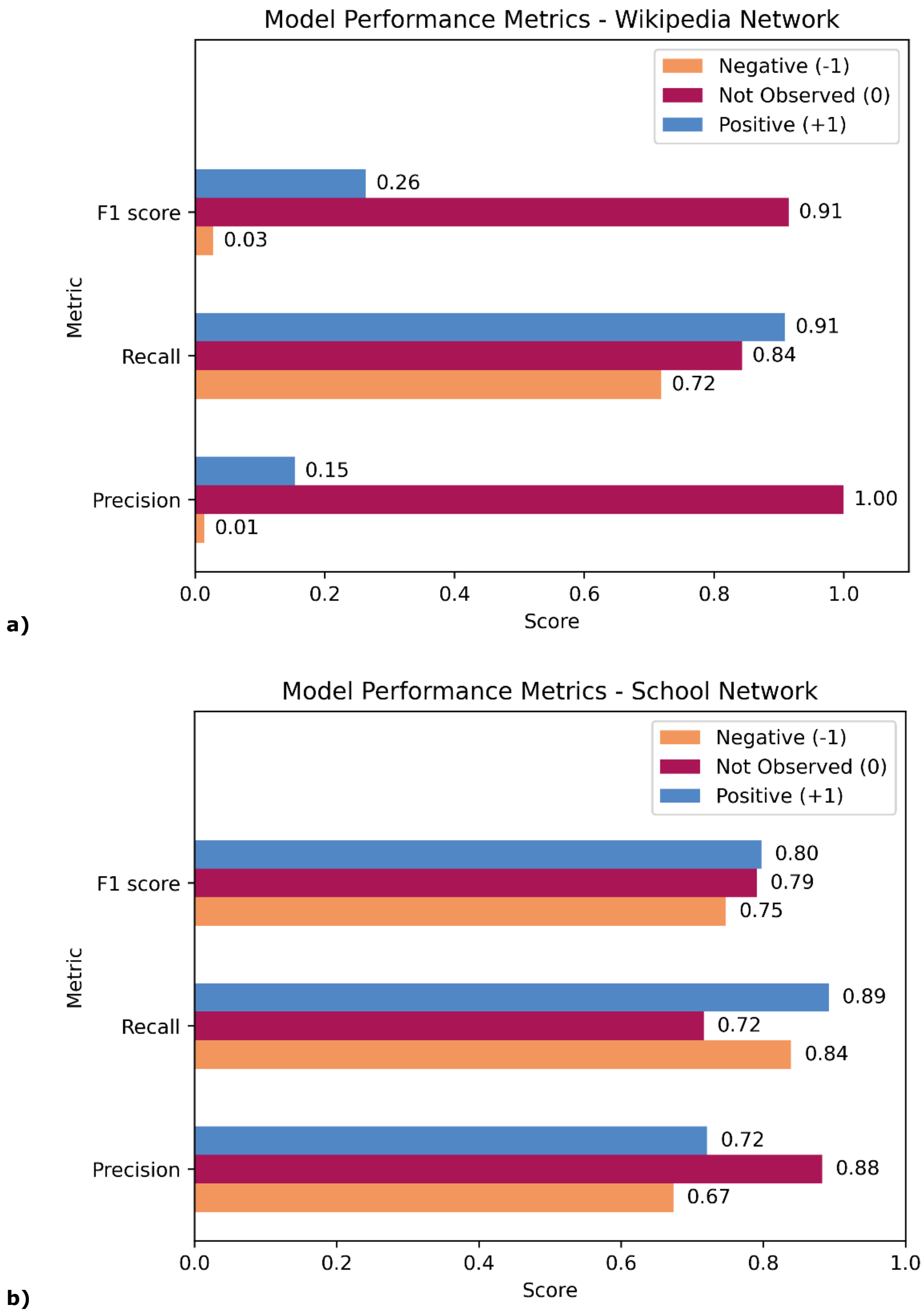


Figure 5: Model Performance Metrics plot for the final Wikipedia Network (a) and School Network (b). The three metrics that are demonstrated are the F1 score, recall and precision (orange = negative, red = not observed and blue = positive).

Precision-Recall Curve

The precision-recall (PR) curves, using the one versus rest (OvR) strategy, for the three classes are shown in Figure 6. Figure 6a shows that the model in the *Wikipedia network* achieves constant precision values over different recall values for the not observed class. This indicates that the performance is excellent. The lines for the positive and negative classes show that precision decreases as recall increases, suggesting that the model is not correctly classifying instances. However, the performance for the positive class is better than the negative class. Figure 6b shows that in the *school network* the model performs overall well for all the classes, with the best performance for the not observed class and the worst for the negative class. For the Wikipedia network, increasing the recall leads to a significant decrease in precision compared to the school network, except for the not observed class. This suggests that as the recall increases, the model captures many false positives leading to a decrease in precision. Therefore, the school network performs better for the model than the Wikipedia network. The reasons for this are mentioned above.

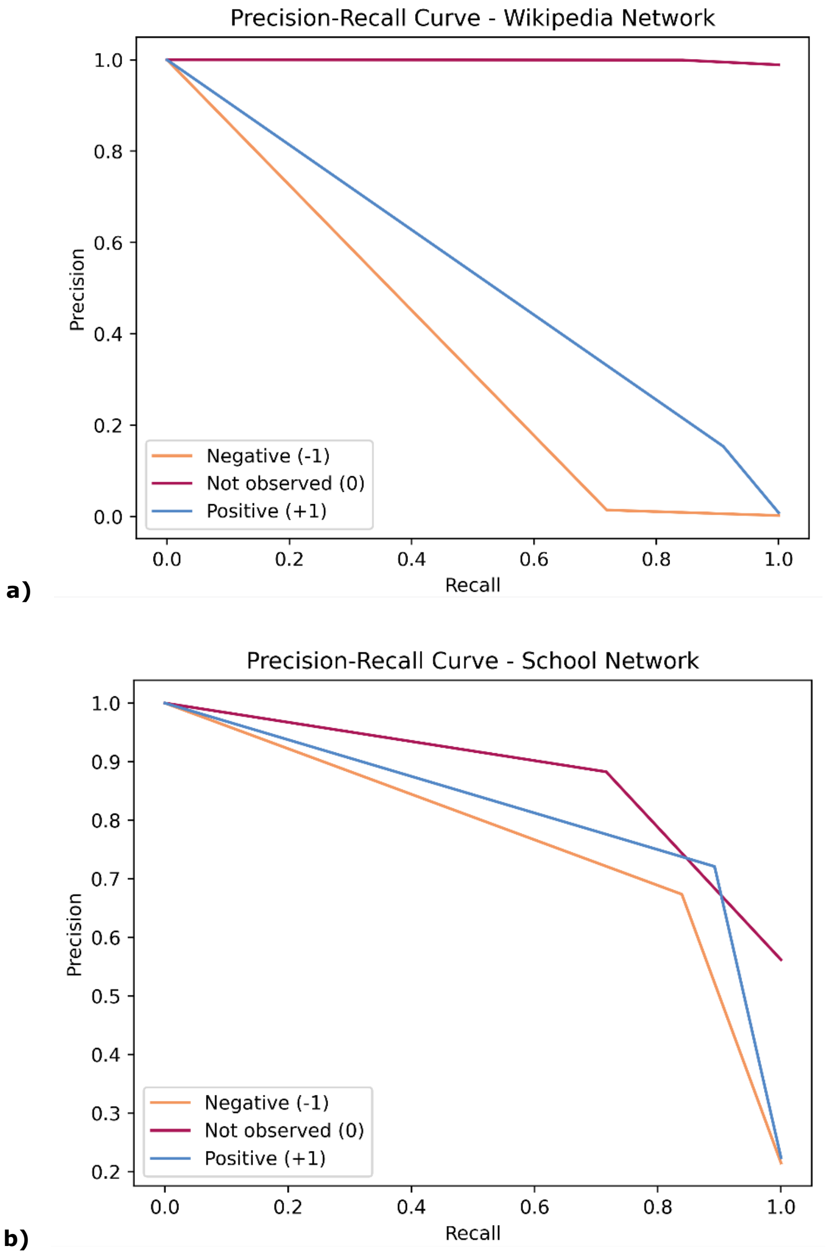
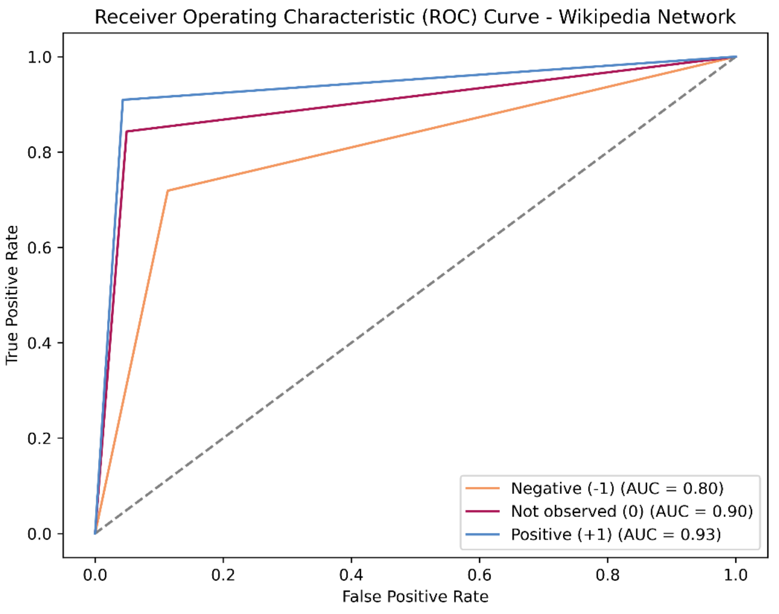


Figure 6: PR Curve for the final Wikipedia Network (a) and School Network (b). This represents the trade-off between the precision (y-axis) and recall (x-axis)for different classification thresholds (orange = negative, red = not observed and blue = positive).

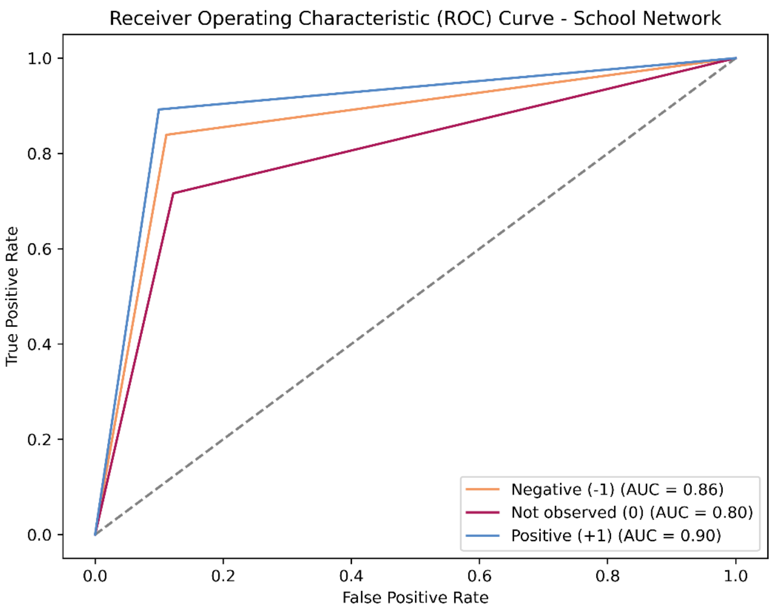
Receiver Operating Characteristic Curve

The Receiver Operating Characteristic (ROC) curve, using the OvR strategy, for the three classes is shown in Figure 7. Figure 7a shows that for the *Wikipedia network* the model performs consistently across all classes. The positive class performs the best (AUC = 0.93) and the negative class performs the worst (AUC = 0.80). This means that the model discriminates strongly between positive and negative instances. Research by Derr TS [5] gives AUC = 0.77 for an online network using only similarity-based metrics and AUC = 0.69 for an online network using path-based metrics. Other research by Robert West et al. [23] gives an AUC of 0.82 for the Wikipedia network using only social theory metrics. The articles show roughly equivalent results; however, our model performs slightly better. This may be due to the inclusion of feature selection, which resulted in the best possible feature set. Figure 7b shows consistent results for the *school network*. The positive class again performs best (AUC = 0.90), however the not observed class performs worst (AUC = 0.80). Research by Feng et al. [17] shows different results for a low clustered real world network (AUC = 0.67) and similar results for a high clustered network (AUC = 0.96). A likely reason for the difference is that Feng et al. only included similarity-based metrics, while our model included a variety of features.

Comparing the two networks, the Wikipedia network has slightly better results (higher AUC) compared to the school network. These results are not consistent with previous results, as the ROC curve is less sensitive to imbalanced datasets and therefore appears more favourable. On the other hand, the PR curve is more reliable because it shows how the classifiers are affected by imbalanced data [42].



a)

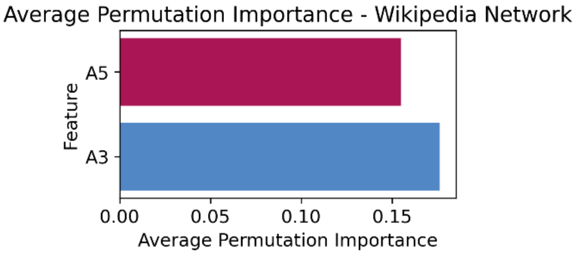


b)

Figure 7: ROC curve for the final Wikipedia Network (a) and School Network (b). This represents model's ability to distinguish between false positives (x-axis) and true positives (y-axis) instances across different thresholds (orange = negative, red = not observed and blue = positive).

Average Permutation Importance

The average permutation importance of the Wikipedia and school networks is shown in Figure 8. A^3 and A^5 are the most notable features for both networks. They represent the paths of length 3 and 5 respectively from each node to other nodes, taking into account the voting sign. Both features are the most informative for predicting the ties in the networks. A^3 may be more important than A^5 because A^5 has some overlapping information with A^3. This could also be the reason why A^4 is less important. Furthermore, in Figure 8b for the *school network,* the three least significant features are respectively outdegree source, eigenvector source and socialstatus target. They capture enough information to be included in the model as suggested by the feature selection but they are not as discriminative as the other features. One reason why many node-level and network-level metrics are selected for the school network may be due to bias in feature selection due to multicollinearity problems, as discussed in Chapter 3.2. This may result in redundant features being selected, leading to an over-emphasis of the results.



a)

b)

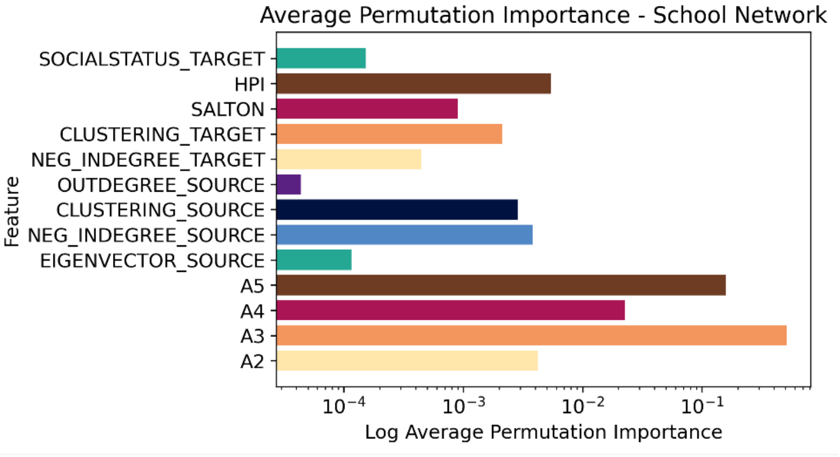


Figure 8: Average Permutation Importance plot for the final Wikipedia Network (a) and School Network (b). This represents the (log) average permutation importance after removing a feature in the feature set.

Limitations

There are also some limitations to this research. First, the final datasets of both networks are large; the Wikipedia dataset has 21402442 ties and the school dataset has 87850 ties. This leads to improved performance, but also to computational bottlenecks. Therefore, the analysis of the Wikipedia network is performed on a subset of the network and comprehensive models (MLR and LGBM) are compared. Furthermore, this research lacks reproducibility as the random seed was inadvertently not set during hyperparameter tuning and feature selection. However, the results can be reproduced by running the Python code (see Appendix V).[4] While the results obtained still reflect the networks in this study, re-running the analysis may yield slightly different results. In addition, this study involves complex concepts that may not have been handled completely accurately. For example, dealing with reciprocal ties by changing the network to undirected results in a loss of information and the social balance metrics are difficult to calculate. Derr TS [5] suggests alternatives for dealing with these complex concepts, but due to time constraints these are not included in this research. Fourth, Wang P [16] discusses the network completion problem, which is also addressed in this research. The collected network data is often incomplete and partially unobserved. Finally, the most important limitation is tackling the imbalance problem which hinders the effectiveness of various link prediction methods [16]. In this research, SKCV and class weights are used as strategies to account for the imbalance; however, these measures were found to be insufficient.

[4] bostaals/networkanalysis: Using machine learning models to predict positive and negative ties in online versus real world networks. (github.com)

5 Conclusion

Network analysis is a powerful tool for exploring network structures from different disciplines. This research explores the world of signed networks, where positive and negative ties coexist. Furthermore, this research compares important metrics for an online network (represented by the Wikipedia network) with a real-world network (represented by the school network). This leads to the following research question: "*To what extent can machine learning models predict positive and negative ties in online and real-world networks?*" To predict positive and negative ties, the data is fed into a feature-based classification model. The feature set consists of three types of ties: negative, not observed and positive. Model selection, hyperparameter tuning and feature selection are performed on a subset of the data to obtain the best setting for the final classification model.

The results show that the school network outperforms the Wikipedia network in terms of F1, recall and precision for all classes except the not observed class. This difference may be due to factors such as the imbalanced datasets, different network structures and different feature sets. However, the analysis showed that paths of length 3 and 5 (A^3 and A^5) are the most significant features for both networks. In particular, the Wikipedia network has a bias towards predicting the not observed class, whereas the school network has a consistently good performance across all classes, as evidenced by the performance metric scores and the PR curve. The ROC curve shows conflicting results, however this plot is less sensitive to imbalanced datasets and therefore unreliable.

In conclusion, the results show that the real world has a higher performance compared to the online network. Due to the high precision in the school network, decisions about seating could be made to improve interactions. Furthermore, the high recall enables effective detection of social isolation and bullying. However, it should be noted that comparing online and real-world networks is challenging due to the different network structure. Therefore, potential future decisions and interventions based on the research should be approached with caution and careful consideration. Investigating different models (e.g., probabilistic graph models and latent-feature methods) and features (e.g., gender) would shed future light on predicting the negative and positive ties in a network. In addition, more similar networks could be compared to provide more generalisable results.

References

[1] Karimi H, Torphy KT, Derr T, Frank KA & Tang J. *Characterizing Teacher Connections in Online Social Media: A Case Study on Pinterest.* Proceedings of the Seventh ACM Conference on Learning@ Scale. 2020 August 12 – 14; Virtual Event (USA). P 249 – 252.

[2] Neal ZP. *A sign of the times? Weak and strong polarization in the U.S. Congress, 1973–2016.* Social Networks. 2020 January; 60; p 103 – 112.

[3] Adegboye OA & Elfaki F. *Network Analysis of MERS Coronavirus within Households, Communities, and Hospitals to Identify Most Centralized and Super-Spreading in the Arabian Penisula, 2012 to 2016.* Canadian Journal of Infectious Diseases and Medical Microbiology. 2018 January 5; 2018; p 1 – 9.

[4] Borgatti SP, Mehra A, Brass DJ, & Labianca G. *Network Analysis in the Social Sciences*. Science. 2009 April 24; 323(5916); p 892 – 895.

[5] Derr TS. *Network Analysis with Negative Links*. WSDM '20. Proceedings of the 13th International Conference on Web Search and Data Mining. 2020 February 3 – 7; Houston, TX (USA). p 1 – 210.

[6] Leskovec J, Huttenlocher D, Kleinberg J. *Predicting Positive and Negative Links in Online Social Networks*. WWW. Proceedings of the 19th international conference on World wide web. 2010 April 26 – 30; Raleigh, North Carolina (USA). 2020 January 20; p 641 – 650.

[7] Clause A, Moore C, Newman MEJ. *Hierarchical structure and the prediction of missing links in networks.* Nature. 2008 May 1; 453(7191); p 98 – 101.

[8] Dasgupta K, Singh R, Viswanathan B, Chakraborty D, Mukherjea S, Nanavati AA, Joshi A. *Social Ties and their Relevance to Churn in Mobile Telecom Networks*. EBDT '08. Proceedings of the 11th International Conference on Extending Database Technology. 2008 March 25 - 30; Nantes (France). p 668 – 677.

[9] Kaur M & Singh S. *Analyzing negative ties in social networks: A survey*. Egyptian Informatics Journal. 2015 September 26; 17(1); p 21 – 43.

[10] Disney A. *Social network analysis 101: centrality measures explained*. Cambridge Intelligence. [Online]. Available from: Social network analysis: Understanding centrality measures (cambridge-intelligence.com) [Accessed 1th June 2023].

[11] Fairchild G & Fries J. *Social Network: Models, Algorithms and Applications*. The University of Iowa. [Online]. Available from: Lecture3.pdf (uiowa.edu) [Accessed 1th of June 2023].

[12] Liben-Nowell D & Kleinberg J. *The link-prediction problem for social networks*. Journal of the American Society for Information Science and Technology. 2007 March 26; 58(7); p 1019 – 1032.

[13] NcPherson M, Smith-Lovin L, Cook JM*. Birds of a Feather: Homophily in Social Networks*. Annual Review of Sociology. 2001; 27(1); p 415- 455*.*

[14] Lü L, Zhou T. *Link prediction in complex networks: A survey*. Elsevier. 2011 March 15; 390(6); p 1150 – 1170.

[15] Chrol B & Bojanowski M. *Proximity-based Methods for Link Prediction*. [Internet]. Available from: Proximity-based Methods for Link Prediction (r-project.org) [Accessed 2th of June 2023].

[16] Wang P, Xu B, Wu , & Zhou X. *Link prediction in social networks: the state-of-the-art.* Science China Information Sciences. 2015 January; 58(1); p 1 – 38.

[17] Feng X, Zhao JC & Xu K. *Link prediction in complex networks: a clustering perspective.* The European Physical Journal B. 2011 August 19; 85(1); p 1 – 9.

[18] Zhou T, Lü L, Zhang YC. *Predicting missing links via local information*. The European Physical Journal B. 2009 June 1; 71(4); p 623 – 630.

[19] Heider F. *Attitudes and cognitive organization.* The Journal of Psychology. 1946; 21(1); p 107 – 112.

[20] Leskovec J, Huttenlocher D, Kleinberg J. *Signed networks in social media.* CHI '10. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2010 April 10 – 15; Atlanta, Georgia (USA). P 1361 – 1370.

[21] Cartwright D, Gleason TC. *The number of paths and cycles in a digraph.* Psychometrika. 1966 June; 31(2); p 179 – 199.

[22] Henley NM, Horsfall RB, De Soto CB. *Goodness of Figure and social structure.* Psychological Review. 1969; 76(2); p 194 – 204.

[23] West R, Paskov HS, Leskovec J, Potts C. *Exploiting Social Network Structure for Person-to-Person Sentiment Analysis*. Transactions of the Association for Computational Linguistics. 2014 September; 2(1); p 297 – 310.

[24] Chiang K, Natarajan N, Tewari A. *Exploiting Longer Cycles for Link Prediction in Signed Networks.* CIKM '11. Proceedings of the 2011 ACM International Conference on Information and Knowledge Management. 2011 October 24 – 28; Glasgow, Scotland (UK); p 1157 – 1162.

[25] Pujari M & Kanawati R. *Supervised Rank Aggregation Approach for Link prediction in Complex Networks.* Proceedings of the 21st International Conference Companion on World Wide Web - WWW '12 Companion. 2012 April 16 – 20. Lyon (France). P 1189 – 1196.

[26] Kashima H & Abe N. *A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction.* Proceedings of the Sixth International Conference on Data Mining (ICDM ' 06). 2006 December 18 – 22. Hong Kong (China). P 340 – 349.

[27] Zhang M, Chen Y. *Link Prediction Based on Graph Neural Networks*. NIPS '18. 32nd Conference on Neural Information processing Systems. 2018 December 3 – 8; Montréal (Canada). p 5171 – 5181.

[28] Menon AK & Elkan C. *Link Prediction via Matrix Factorization.* Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases. 2011 September 5 – 9. Athens (Greece). P 437 – 452.

[29] Leskovec J. *Wikipedia Requests for Adminship (with text).* [Internet]. Available from: SNAP: Network datasets: Wikipedia Requests for Adminship (with text) (stanford.edu). [Accessed 3rd of June 2023].

[30] Kalter F, Heath AF, Hewstone M, Jonsson JO, Kalmijn M, Kogan I and van Tubergen F. 2016. *Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU) – Reduced version. Reduced data file for download and off-site use.* GESIS Data Archive, Cologne, ZA5656 Data file Version 1.2.0, doi:10.4232/cils4eu.5656.1.2.0.

[31] CILS4EU. 2016. *Children of Immigrants Longitudinal Survey in Four European Countries. Codebook.* Wave 1 – 2010/2011, v1.2.0. Mannheim: Mannheim University.

[32] Kolaczyk ED. *Statistical Analysis of Network Data: Methods and Models (Springer Series in Statistics).* 2009th Edition. New York: Springer Science; 2009.

[33] Chan JT. *Snowball Sampling and Sample Selection in a Social Network.* SSRN. 2015 September 1; p 1 – 23.

[34] Prutsky S, Patnaik S, Dash SK. *SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer.* Frontiers in Nanotechnology. 2022 August 19; 4; p 1 – 12.

[35] Muralidhar KSV. *What is Stratified Cross-Validation in Machine Learning?* [Internet]. Available from: What is Stratified Cross-Validation in Machine Learning? | by KSV Muralidhar | Towards Data Science. [Accessed 14th of June 2023].

[36] Dwivedi R. What is LightGBM Algorithm, How to use it? [Internet]. Available from: What is LightGBM Algorithm, How to use it? | Analytics Steps. [Accessed 23th of June 2023].

[37] Weerts HJP, Müller AC, Vanschoren J. *Importance of Tuning Hyperparameters of Machine Learning Algorithms*. Technische Universiteit Eindhoven. 2020 July 15; p 1 -17.

[38] Beheshti N. *Cross Validation and Grid Search.* [Internet]. Available from: Cross Validation and Grid Search. Using sklearn's GridSearchCV on random… | by Nima Beheshti | Towards Data Science. [Accessed 14th of June 2023]

[39] Bahl A, Hellack B, Balas M, Dinischiotu A, Wiemann M, Brinkmann J, Luch A, Renard BY, Haase A. *Recursive feature elimination in random forest classification supports nanomaterial grouping*. NanoImpact. 2019 March; 15; p 1 – 12.

[40] Harshilsanghvi. *Recursive Feature Elimination with Cross-Validation in Scikit Learn.* [Internet]. Available from: Recursive Feature Elimination with Cross-Validation in Scikit Learn - GeeksforGeeks. [Accessed 14th of June 2023].

[41] Din SMU, Mehmood SK, Shahzad A, Ahmad I, Davidyants A & Abu-Rumman A. *The Impact of Behavioral Biases on Herding Behavior of Investors in Islamic Financial Products.* Frontiers in Psychology. 2021 February 04; 11(2020); p 1 – 10.

[42] Saito T & Rehmsmeier M. *The Precision-Recall Plot is More In Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.* 2015 March 4. PLOS ONE; 10(3); p 1 – 21.

Appendix

I Feature set description

Feature	Explanation	Wikipedia	School
SOURCE	Username / student ID of voter.	X	X
TARGET	Username / student ID of target.	X	X
DAT	The date and time of the vote.	X	
VOTE	The sign of the vote (-1 = negative, 1 = positive).	X	X
CLASSID_{…}	Unique ID for class of student.		X
SCHOOLID_{…}	Unique ID for school of student.		X
A2	Path length, k = 2; Captures information about common neighbours. Represents the social balance theory.	X	X
A3	Path length, k = 3.	X	X
A4	Path length, k = 4.	X	X
A5	Path length, k = 5 .	X	X
INDEGREE_{…}	Number of incoming links.	X	X
OUTDEGREE_{…}	Number of outgoing links.	X	X
POS_INDEGREE_{…}	Number of incoming positive links.	X	X
NEG_INGEDREE_{…}	Number of incoming negative links.	X	X
EIGENVECTOR_{…}	Centrality of a node within a network, based on its connections to other high influential nodes.	X	X
PAGERANK_{…}	Importance score based on the network structure and eigenvector.	X	X
CLUSTERING_{…}	Reflects the level of local cohesion within the network.	X	X
CN	Common Neighbours: the similarity based on the count of shared connections between two nodes, $\|common\ neighbors(u,v)\|\ .$	X	X
SALTON	Salton Cosine Similarity: The similarity based on the cosine of the angle between the neighbour vectors, $\frac{\|common\ neighbors(u,v\|}{\sqrt{\|neighbors(u)\| \cdot \|neighbors(v)}}.$	X	X

JACCARD	Jaccard Coefficient: The similarity based on the ratio of common neighbours to the total neighbours, accounting for overlap, $\frac{\vert common\_neighbors(u,v)}{\vert neighbors(u)\vert + \vert neighbors(v)\vert - \vert common\_neighbors(u,v)}$.	X	X
SORENSEN	Sørenson Index: The similarity based on the size of the intersection to the sum of the individual neighbour set size, $\frac{2 \cdot \vert common\ neighbors(u,v)\vert}{\vert neighbors(u)\vert + \vert neighbors(v)\vert}$.	X	X
HPI	Hub Promoted Index: The similarity based on the extent to which nodes promote the connectivity of other nodes, $\frac{\vert common\ neighbors(u,v)\vert}{\min(\vert neighbors(u)\vert, \vert neighbors(v)\vert}$.	X	X
HDI	Hub Depressed Index: The similarity based on the degree of exclusion of nodes within a network, $\frac{\vert common\ neighbors(u,v)\vert}{\max(\vert neighbors(u)\vert, \vert neighbors(v)\vert}$.	X	X
PA	Preferential Attachment: The tendency of new nodes to preferentially attach to highly connected nodes, $\frac{k_u}{\sum_{v \in V} k_v}$.	X	X
AA	Adamic-Adar Coefficient: The similarity emphasizing the importance of connecting through less common neighbours, $\sum_{w \in common\ neighbors(u,v)} \frac{1}{\log(\vert neighbors(w\vert)}$.	X	X
RA	Resource Allocation: The similarity emphasizing the distribution of information through sparser connections, $\sum_{w \in common\ neighbors(u,v)} \frac{1}{(\vert neighbors(w\vert)}$.	X	X
SS_TIME_{…}	Reflects the social status theory under the assumption that earlier votes have a higher status.	X	

SOCIALSTATUS_{…}	Reflects the social status theory based on research from Derr TS, $\sum (eigenvector \cdot positive\ ties) - (eigenvector \cdot negative\ ties)$.	X	X
SS_POPULARITY_{…}	Reflects the social status theory under the assumption that students with more popularity votes have a higher status.		X

{..} indicates that this feature has one feature for the SOURCE node and one for the TARGET node

II Model performance metric

A confusion matrix sums up the amount of predicted versus true observations. The confusion matrix is computed on the test set. An overview for the confusion matrix of a multinomial classification is in Figure 1.

		Predicted		
		Negative (-1)	Not Observed (0)	Positive (+1)
Observed	Negative (-1)	TP_-1	FP_-1	FP_-1
	Not Observed (0)	FP_0	TP_0	FP_0
	Positive (+1)	FP_1	FP_1	TP_1

TP = True Positive
FP = False Positive

The precision, recall and F1 of each class i can be calculated using the following formula's respectively using the One-Versus-Rest (OVR) strategy,

$$\boldsymbol{Precision_i} = \frac{TP_i}{TP_i + FP_i} \text{ ,}$$

$$\boldsymbol{Recall_i} = \frac{TP_i}{TP_i + FN_i}, where\ FN_i = Total_i - TP_i - FP_i \text{ ,}$$

$$\boldsymbol{F1}_i = 2 * \frac{Precision_i * Recall_i}{Precision_i + Recall_i} \text{ .}$$

Moreover, the macro F1 has been measured to deal with class imbalance. The formula for the macro F1 is,

$$\boldsymbol{Macro\ F1} = \frac{F1_{-1} + F1_0 + F1_1}{3} \text{ .}$$

Finally, the AUC is calculated using the same OVR strategy. The True Positive Rate (TPR) and False Positive Rate (FPR) are plotted, and the AUC is obtained.

III Full data exploration results

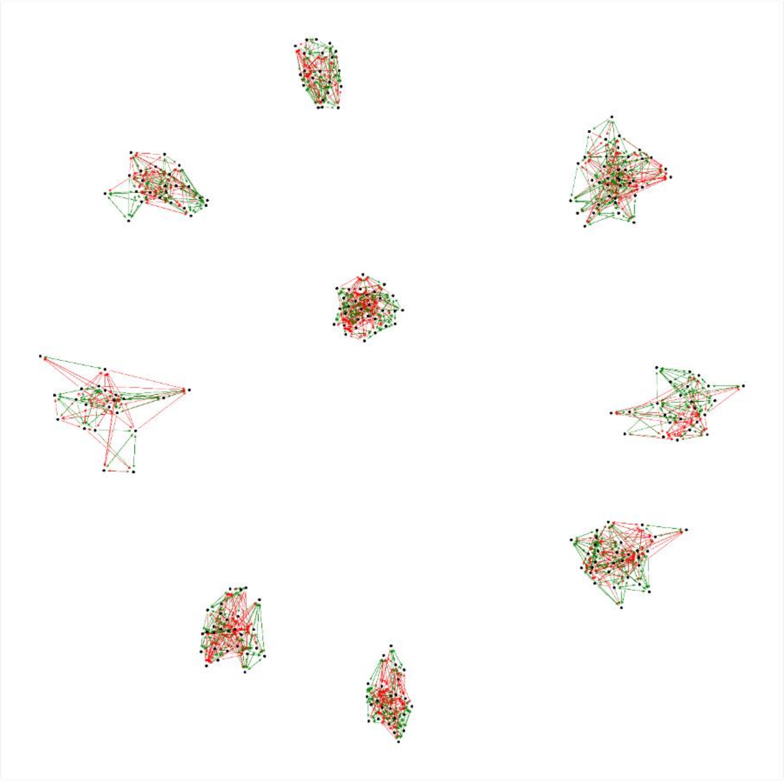


Figure 1: Visualization of the subset of the school network.

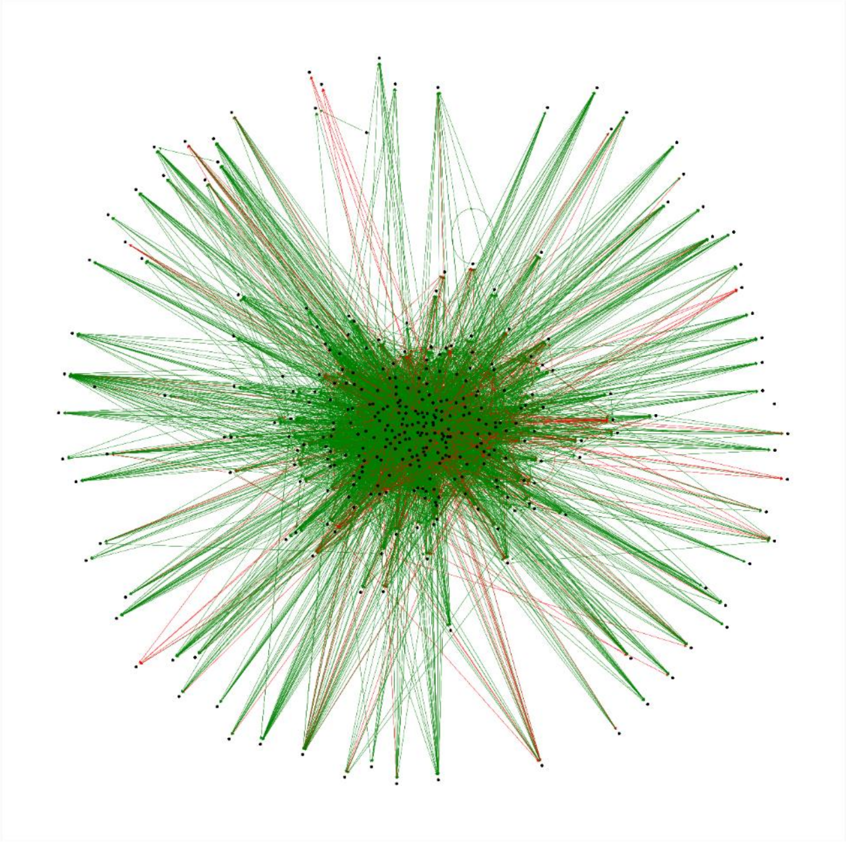
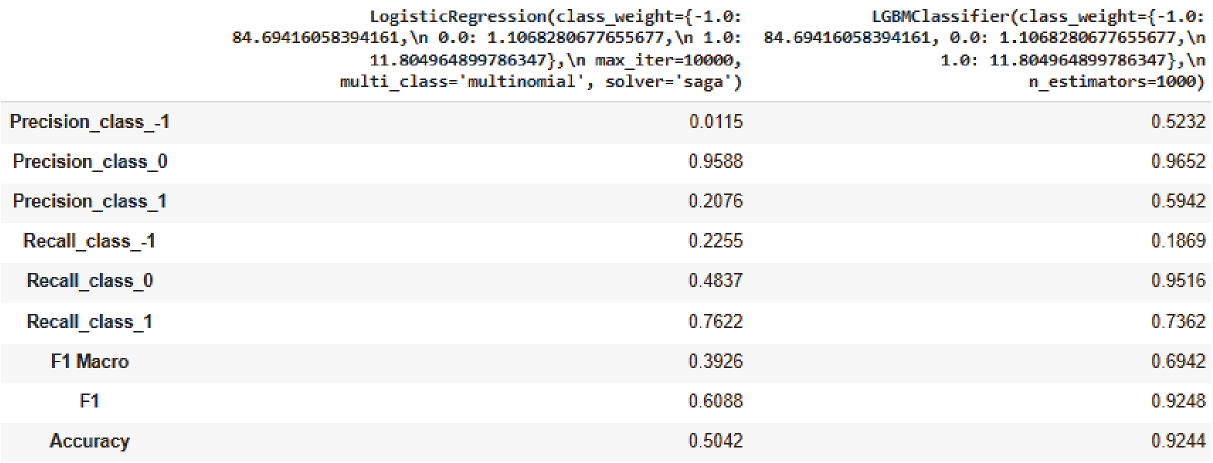


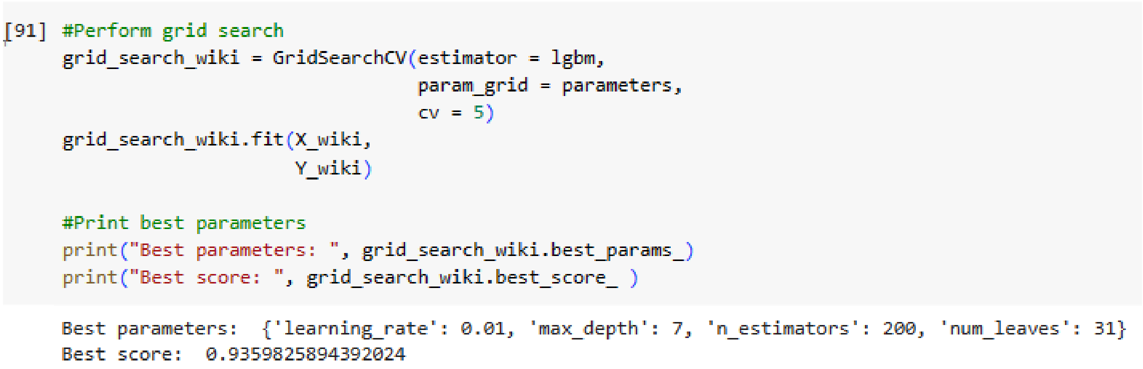
Figure 2: Visualization of the subset of the Wikipedia network.

IV Full analysis results

1. Wikipedia subset

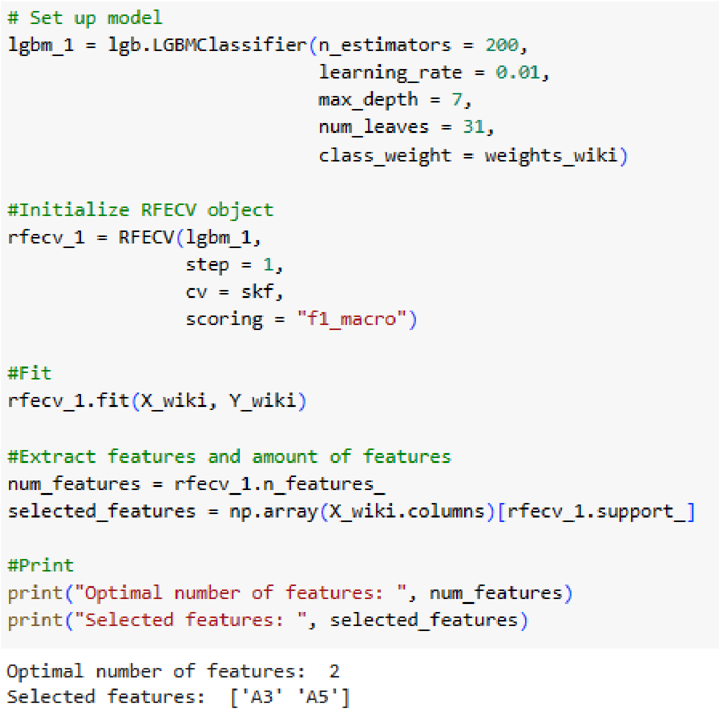


	LogisticRegression(class_weight={-1.0: 84.69416058394161,\n 0.0: 1.1068280677655677,\n 1.0: 11.804964899786347},\n max_iter=10000, multi_class='multinomial', solver='saga')	LGBMClassifier(class_weight={-1.0: 84.69416058394161, 0.0: 1.1068280677655677,\n 1.0: 11.804964899786347},\n n_estimators=1000)
Precision_class_-1	0.0115	0.5232
Precision_class_0	0.9588	0.9652
Precision_class_1	0.2076	0.5942
Recall_class_-1	0.2255	0.1869
Recall_class_0	0.4837	0.9516
Recall_class_1	0.7622	0.7362
F1 Macro	0.3926	0.6942
F1	0.6088	0.9248
Accuracy	0.5042	0.9244



```
[91] #Perform grid search
     grid_search_wiki = GridSearchCV(estimator = lgbm,
                                     param_grid = parameters,
                                     cv = 5)
     grid_search_wiki.fit(X_wiki,
                          Y_wiki)

     #Print best parameters
     print("Best parameters: ", grid_search_wiki.best_params_)
     print("Best score: ", grid_search_wiki.best_score_ )
```

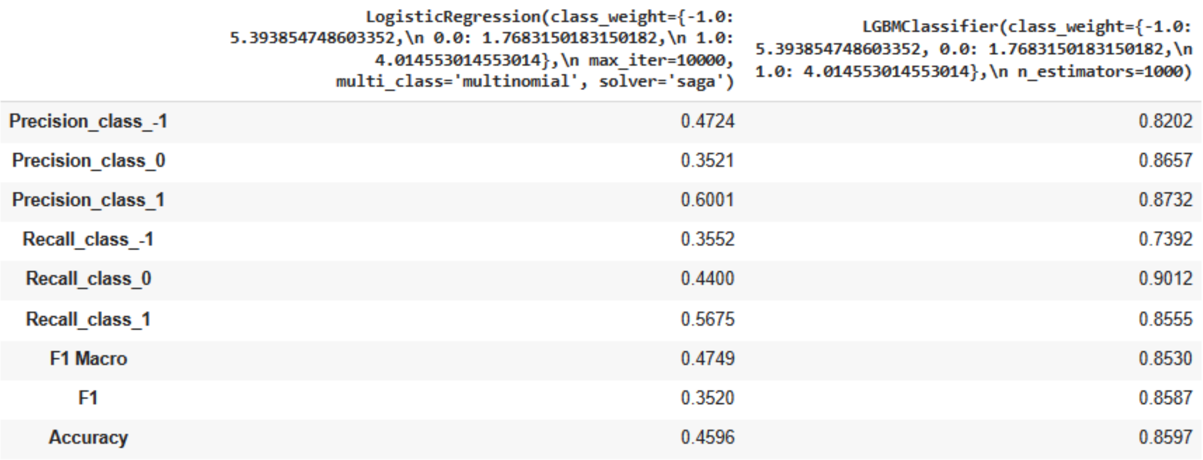
```
     Best parameters:  {'learning_rate': 0.01, 'max_depth': 7, 'n_estimators': 200, 'num_leaves': 31}
     Best score:  0.9359825894392024
```



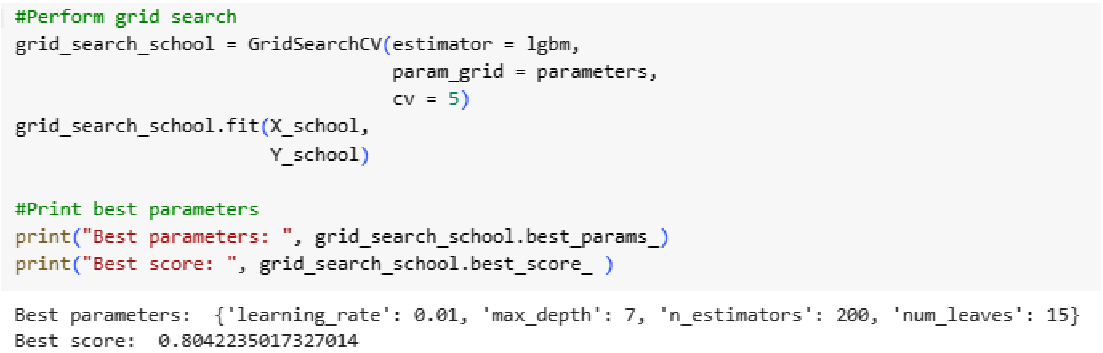
```
# Set up model
lgbm_1 = lgb.LGBMClassifier(n_estimators = 200,
                            learning_rate = 0.01,
                            max_depth = 7,
                            num_leaves = 31,
                            class_weight = weights_wiki)

#Initialize RFECV object
rfecv_1 = RFECV(lgbm_1,
                step = 1,
                cv = skf,
                scoring = "f1_macro")

#Fit
rfecv_1.fit(X_wiki, Y_wiki)

#Extract features and amount of features
num_features = rfecv_1.n_features_
selected_features = np.array(X_wiki.columns)[rfecv_1.support_]

#Print
print("Optimal number of features: ", num_features)
print("Selected features: ", selected_features)
```

```
Optimal number of features:  2
Selected features:  ['A3' 'A5']
```

2. School subset

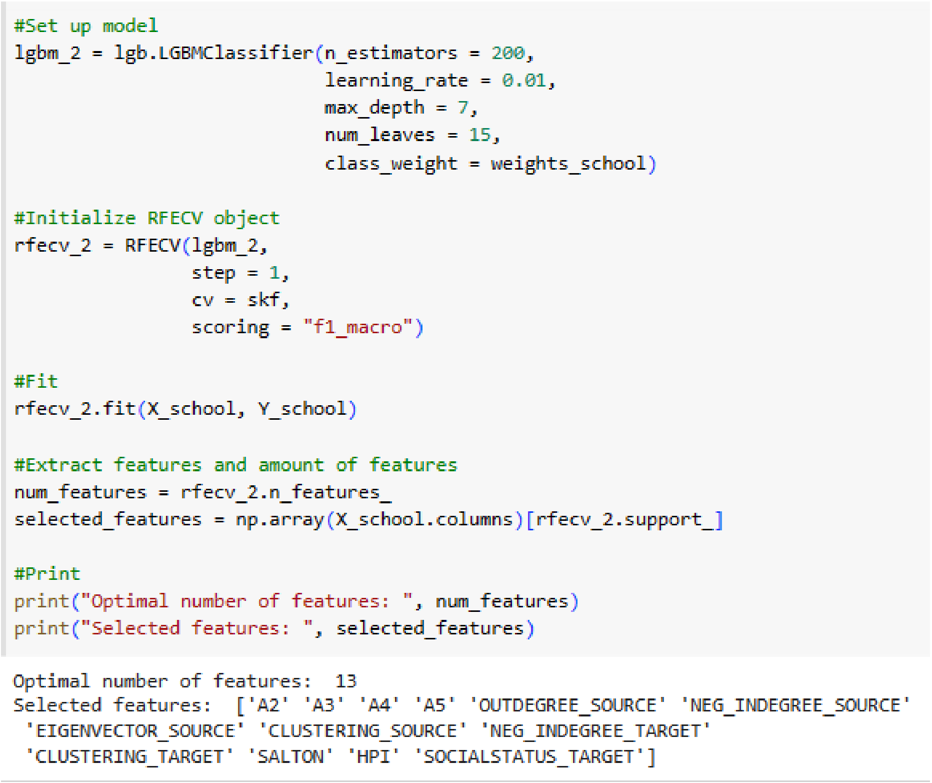


	LogisticRegression(class_weight={-1.0: 5.393854748603352,\n 0.0: 1.7683150183150182,\n 1.0: 4.014553014553014},\n max_iter=10000, multi_class='multinomial', solver='saga')	LGBMClassifier(class_weight={-1.0: 5.393854748603352, 0.0: 1.7683150183150182,\n 1.0: 4.014553014553014},\n n_estimators=1000)
Precision_class_-1	0.4724	0.8202
Precision_class_0	0.3521	0.8657
Precision_class_1	0.6001	0.8732
Recall_class_-1	0.3552	0.7392
Recall_class_0	0.4400	0.9012
Recall_class_1	0.5675	0.8555
F1 Macro	0.4749	0.8530
F1	0.3520	0.8587
Accuracy	0.4596	0.8597



```
#Perform grid search
grid_search_school = GridSearchCV(estimator = lgbm,
                                  param_grid = parameters,
                                  cv = 5)
grid_search_school.fit(X_school,
                       Y_school)

#Print best parameters
print("Best parameters: ", grid_search_school.best_params_)
print("Best score: ", grid_search_school.best_score_ )
```

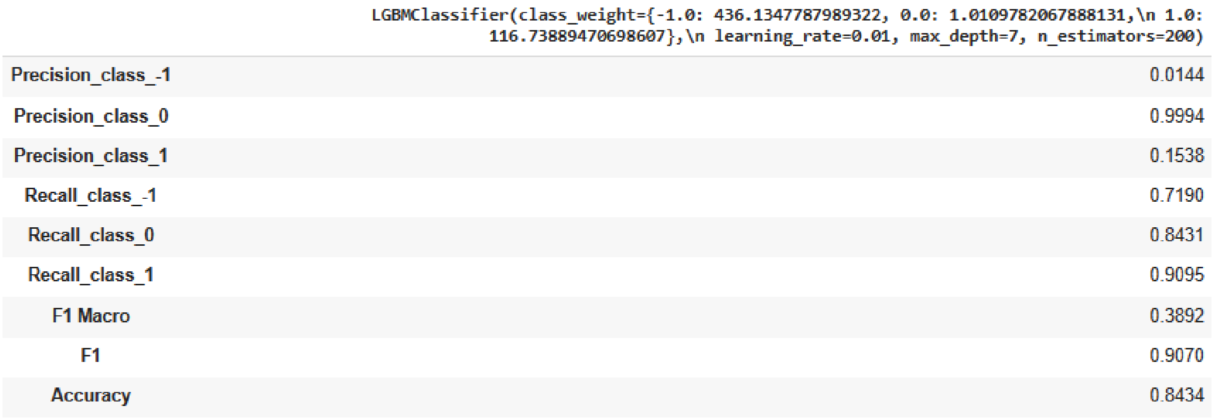
```
Best parameters:  {'learning_rate': 0.01, 'max_depth': 7, 'n_estimators': 200, 'num_leaves': 15}
Best score:  0.8042235017327014
```



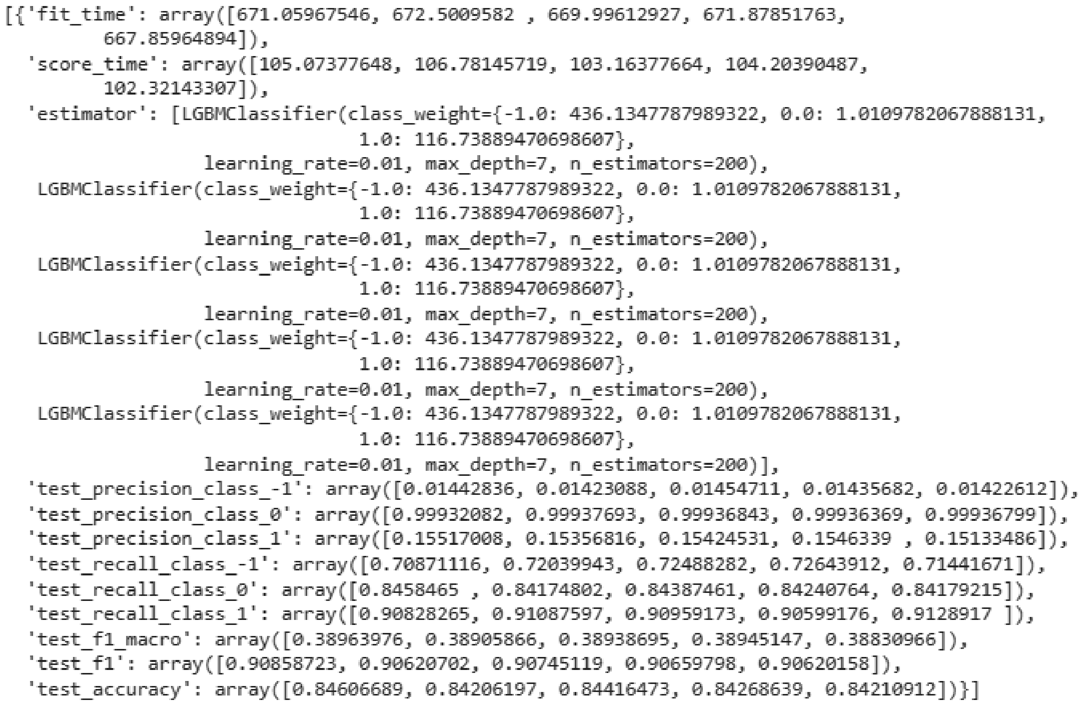
```
#Set up model
lgbm_2 = lgb.LGBMClassifier(n_estimators = 200,
                            learning_rate = 0.01,
                            max_depth = 7,
                            num_leaves = 15,
                            class_weight = weights_school)

#Initialize RFECV object
rfecv_2 = RFECV(lgbm_2,
                step = 1,
                cv = skf,
                scoring = "f1_macro")

#Fit
rfecv_2.fit(X_school, Y_school)

#Extract features and amount of features
num_features = rfecv_2.n_features_
selected_features = np.array(X_school.columns)[rfecv_2.support_]

#Print
print("Optimal number of features: ", num_features)
print("Selected features: ", selected_features)
```

```
Optimal number of features:  13
Selected features:  ['A2' 'A3' 'A4' 'A5' 'OUTDEGREE_SOURCE' 'NEG_INDEGREE_SOURCE'
 'EIGENVECTOR_SOURCE' 'CLUSTERING_SOURCE' 'NEG_INDEGREE_TARGET'
 'CLUSTERING_TARGET' 'SALTON' 'HPI' 'SOCIALSTATUS_TARGET']
```

3. Wikipedia full network

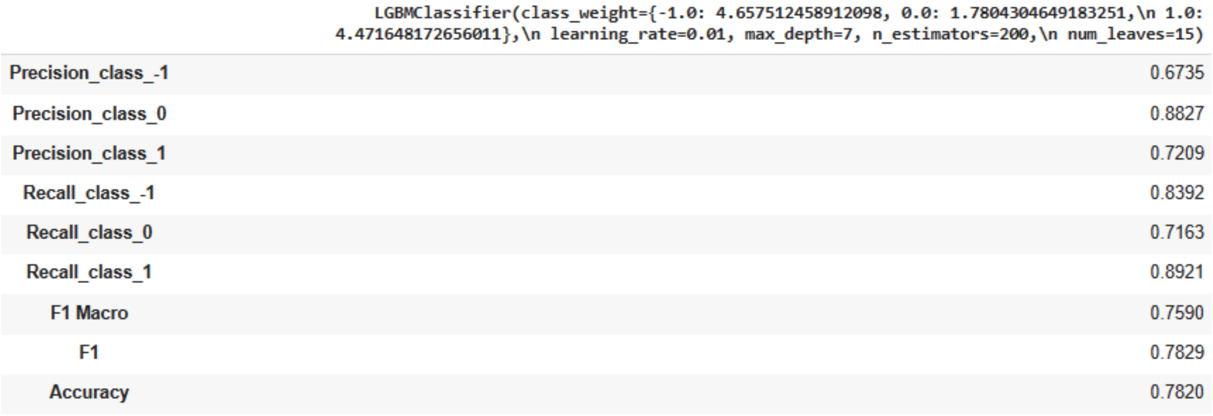


	LGBMClassifier(class_weight={-1.0: 436.1347787989322, 0.0: 1.0109782067888131,\n 1.0: 116.73889470698607},\n learning_rate=0.01, max_depth=7, n_estimators=200)
Precision_class_-1	0.0144
Precision_class_0	0.9994
Precision_class_1	0.1538
Recall_class_-1	0.7190
Recall_class_0	0.8431
Recall_class_1	0.9095
F1 Macro	0.3892
F1	0.9070
Accuracy	0.8434

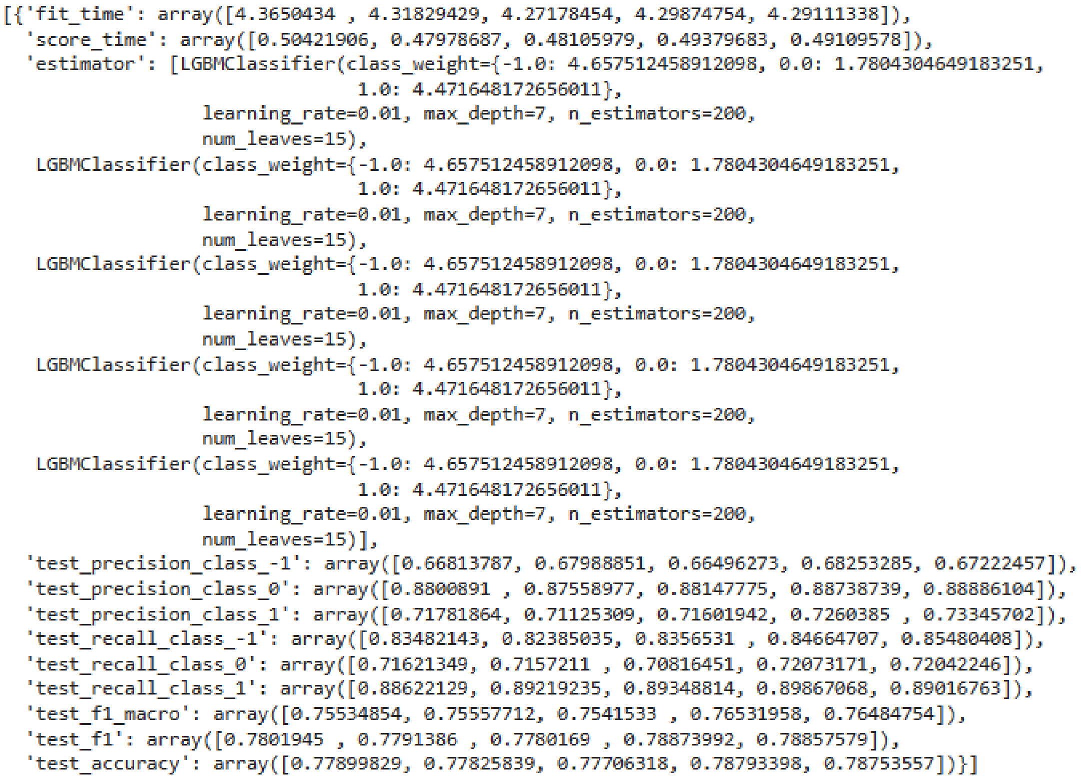


```
[{'fit_time': array([671.05967546, 672.5009582 , 669.99612927, 671.87851763,
         667.85964894]),
  'score_time': array([105.07377648, 106.78145719, 103.16377664, 104.20390487,
         102.32143307]),
  'estimator': [LGBMClassifier(class_weight={-1.0: 436.1347787989322, 0.0: 1.0109782067888131,
                              1.0: 116.73889470698607},
                learning_rate=0.01, max_depth=7, n_estimators=200),
   LGBMClassifier(class_weight={-1.0: 436.1347787989322, 0.0: 1.0109782067888131,
                              1.0: 116.73889470698607},
                learning_rate=0.01, max_depth=7, n_estimators=200),
   LGBMClassifier(class_weight={-1.0: 436.1347787989322, 0.0: 1.0109782067888131,
                              1.0: 116.73889470698607},
                learning_rate=0.01, max_depth=7, n_estimators=200),
   LGBMClassifier(class_weight={-1.0: 436.1347787989322, 0.0: 1.0109782067888131,
                              1.0: 116.73889470698607},
                learning_rate=0.01, max_depth=7, n_estimators=200),
   LGBMClassifier(class_weight={-1.0: 436.1347787989322, 0.0: 1.0109782067888131,
                              1.0: 116.73889470698607},
                learning_rate=0.01, max_depth=7, n_estimators=200)],
  'test_precision_class_-1': array([0.01442836, 0.01423088, 0.01454711, 0.01435682, 0.01422612]),
  'test_precision_class_0': array([0.99932082, 0.99937693, 0.99936843, 0.99936369, 0.99936799]),
  'test_precision_class_1': array([0.15517008, 0.15356816, 0.15424531, 0.1546339 , 0.15133486]),
  'test_recall_class_-1': array([0.70871116, 0.72039943, 0.72488282, 0.72643912, 0.71441671]),
  'test_recall_class_0': array([0.8458465 , 0.84174802, 0.84387461, 0.84240764, 0.84179215]),
  'test_recall_class_1': array([0.90828265, 0.91087597, 0.90959173, 0.90599176, 0.9128917 ]),
  'test_f1_macro': array([0.38963976, 0.38905866, 0.38938695, 0.38945147, 0.38830966]),
  'test_f1': array([0.90858723, 0.90620702, 0.90745119, 0.90659798, 0.90620158]),
  'test_accuracy': array([0.84606689, 0.84206197, 0.84416473, 0.84268639, 0.84210912])}]
```

4. School full network



	LGBMClassifier(class_weight={-1.0: 4.657512458912098, 0.0: 1.7804304649183251,\n 1.0: 4.471648172656011},\n learning_rate=0.01, max_depth=7, n_estimators=200,\n num_leaves=15)
Precision_class_-1	0.6735
Precision_class_0	0.8827
Precision_class_1	0.7209
Recall_class_-1	0.8392
Recall_class_0	0.7163
Recall_class_1	0.8921
F1 Macro	0.7590
F1	0.7829
Accuracy	0.7820



```
[{'fit_time': array([4.3650434 , 4.31829429, 4.27178454, 4.29874754, 4.29111338]),
  'score_time': array([0.50421906, 0.47978687, 0.48105979, 0.49379683, 0.49109578]),
  'estimator': [LGBMClassifier(class_weight={-1.0: 4.657512458912098, 0.0: 1.7804304649183251,
                              1.0: 4.471648172656011},
                learning_rate=0.01, max_depth=7, n_estimators=200,
                num_leaves=15),
   LGBMClassifier(class_weight={-1.0: 4.657512458912098, 0.0: 1.7804304649183251,
                              1.0: 4.471648172656011},
                learning_rate=0.01, max_depth=7, n_estimators=200,
                num_leaves=15),
   LGBMClassifier(class_weight={-1.0: 4.657512458912098, 0.0: 1.7804304649183251,
                              1.0: 4.471648172656011},
                learning_rate=0.01, max_depth=7, n_estimators=200,
                num_leaves=15),
   LGBMClassifier(class_weight={-1.0: 4.657512458912098, 0.0: 1.7804304649183251,
                              1.0: 4.471648172656011},
                learning_rate=0.01, max_depth=7, n_estimators=200,
                num_leaves=15),
   LGBMClassifier(class_weight={-1.0: 4.657512458912098, 0.0: 1.7804304649183251,
                              1.0: 4.471648172656011},
                learning_rate=0.01, max_depth=7, n_estimators=200,
                num_leaves=15)],
  'test_precision_class_-1': array([0.66813787, 0.67988851, 0.66496273, 0.68253285, 0.67222457]),
  'test_precision_class_0': array([0.8800891 , 0.87558977, 0.88147775, 0.88738739, 0.88886104]),
  'test_precision_class_1': array([0.71781864, 0.71125309, 0.71601942, 0.7260385 , 0.73345702]),
  'test_recall_class_-1': array([0.83482143, 0.82385035, 0.8356531 , 0.84664707, 0.85480408]),
  'test_recall_class_0': array([0.71621349, 0.7157211 , 0.70816451, 0.72073171, 0.72042246]),
  'test_recall_class_1': array([0.88622129, 0.89219235, 0.89348814, 0.89867068, 0.89016763]),
  'test_f1_macro': array([0.75534854, 0.75557712, 0.7541533 , 0.76531958, 0.76484754]),
  'test_f1': array([0.7801945 , 0.7791386 , 0.7780169 , 0.78873992, 0.78857579]),
  'test_accuracy': array([0.77899829, 0.77825839, 0.77706318, 0.78793398, 0.78753557])}]
```

5. Confusion Matrix

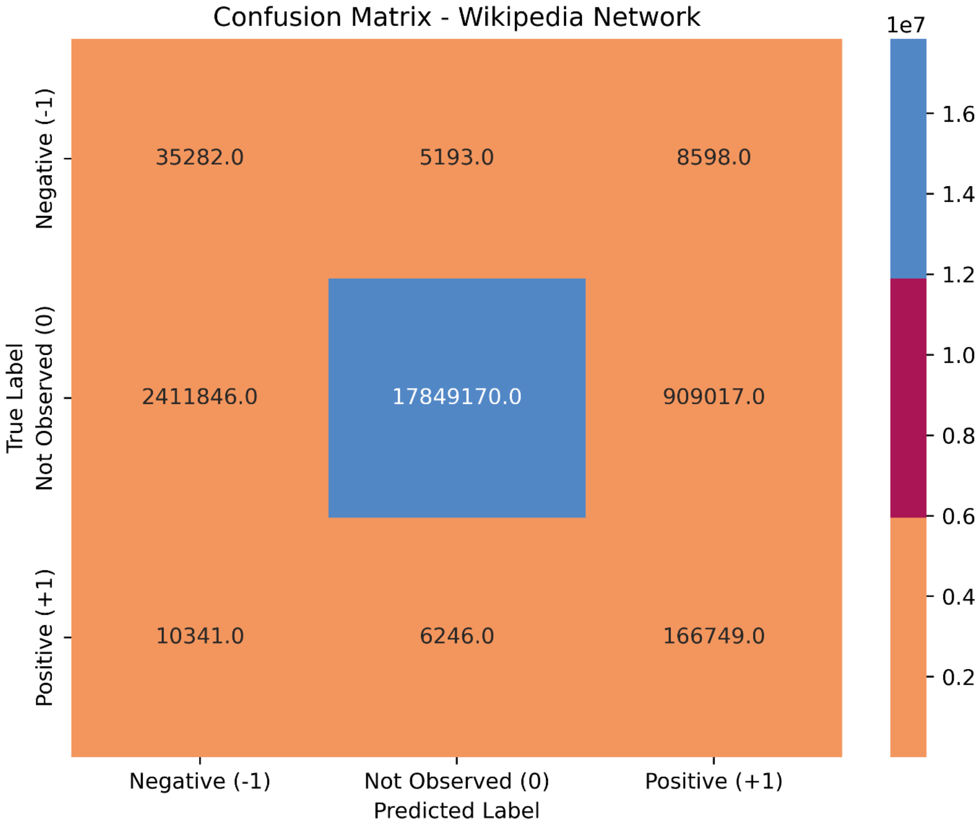


Figure 1: Confusion Matrix for the full wikipedia network.

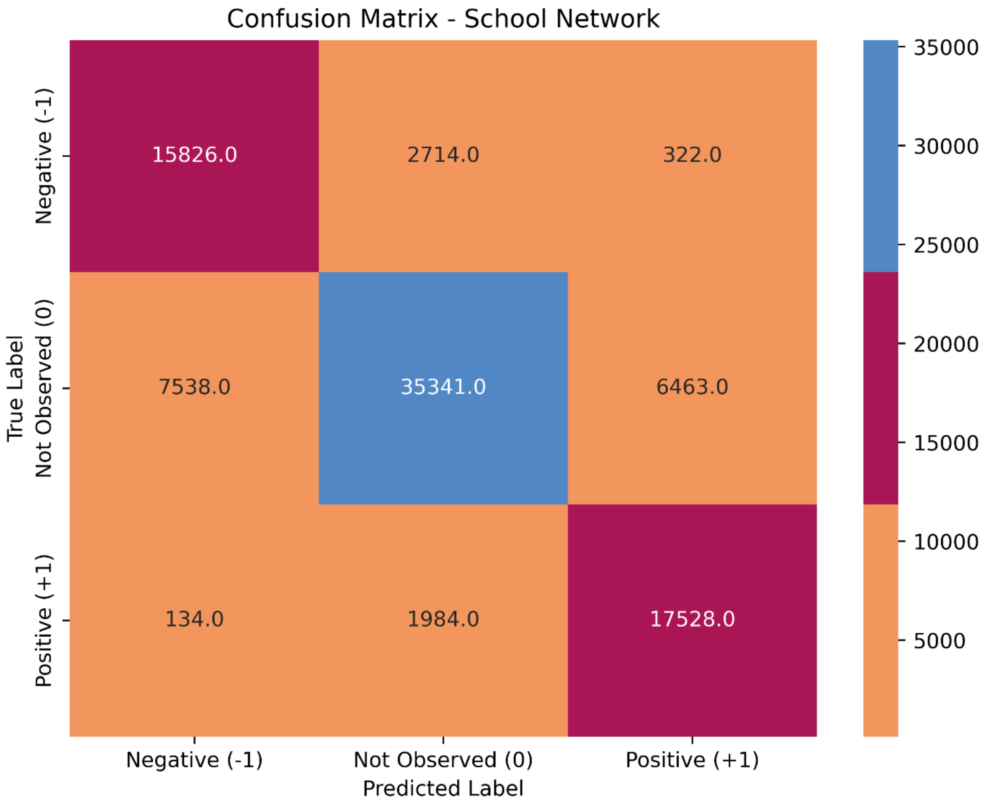


Figure 2: Confusion matrix for the full school network.

V Annotated scripts of analysis and method settings

Link to github:

bostaals/networkanalysis: Using machine learning models to predict positive and negative ties in online versus real world networks. (github.com)