Web of Students: Class-Level Friendship Network Discovery from Educational Big Data

Teng Guo[1], Tao Tang[2], Dongyu Zhang[1], Jianxin Li[3], and Feng Xia[2(✉)]

[1] School of Software, Dalian University of Technology, Dalian 116620, China
[2] School of Engineering, IT and Physical Sciences, Federation University Australia, Ballarat, VIC 3353, Australia
f.xia@ieee.org
[3] School of Information Technology, Deakin University, Melbourne, VIC 3125, Australia

Abstract. Classmate friendships are a major aspect of university social experience. Taking classes together is one of the main ways for students to build friendships. Consequently, class-level friendship networks have attracted tremendous attention from researchers. They are also very useful in student support and early intervention. However, these networks are normally invisible for educators. Discovering such an important web of students effectively is a pressing problem. Against this background, we propose a data-driven framework called CANDY which automatically discovers the class-level friendship networks based on educational big data. We first represent features through representation learning methods. Secondly, the data is augmented with the randomly shuffling method. Thirdly, a conditional generative adversarial network model is used to mine the class-level friendship networks. A deep adversarial optimization strategy is proposed here for problems caused by network sparsity. To evaluate the performance of the proposed approach, we build a real-world dataset that contains rich student information. Extensive experiments have been conducted and the results demonstrate the effectiveness of our framework.

Keywords: Social network analysis · Educational big data · Generative adversarial networks · Friendship networks

1 Introduction

Friendship plays an important role in everyone's life. Whether in personal development or social well-being, friendship matters a lot for everyone, especially for university students who are not physically and mentally matured [19,25]. Taking classes together is one of the main ways for students to build friendships. Consequently, class-level friendship networks have attracted researchers tremendous attention [18,23]. Studies indicate that abnormal behaviors including drinking, stress, depression, and suicide in various age of student groups are related to their repugnant friendships [20,21]. However, such an important web of students is invisible and difficult to discover, which poses a challenge for the education management department. In this case, a significant topic of educational research is the discovery of class-level friendship networks based on known student information data stored in the education management system.

W. Zhang et al. (Eds.): WISE 2021, LNCS 13080, pp. 497–511, 2021.
https://doi.org/10.1007/978-3-030-90888-1_38

Discovering friendship networks among classmates faces tremendous challenges because their social choices might be impacted by various factors. Previous research shows that the social choice of students could be impacted by appearance features, psychology features, intellectual features, behavioral features, and various kinds of similar features as well [15,32,35]. From a methodological point of view, current research in this field can mainly be divided into two categories: questionnaire-based research [35] and link prediction-based research [22]. The questionnaire method is frequently used by statisticians to analyze the relationship between friendship and demographic characteristics through small batches of samples. In terms of efficiency, this method is costly and time-consuming, and difficult to use as a means of daily management. The link prediction aims to predict latent relationships based on known friendships. In a word, neither of these approaches can meet the requirements of practical applications that can infer the web of students automatically based on the data stored in the educational management system to assist the daily management of the university.

With the rise of artificial intelligence in this technological era [8,11,12], the research paradigm turns into the fourth stage. Big data technology has greatly advanced network science, which enriches the means for social network analysis [10,34,36]. These advancements provide us with an unprecedented opportunity to reveal the laws behind the web of students. Nevertheless, new challenges and limitations are introduced as well. Friendship relationships between people are interrelated rather than independent. Such a high-order network feature results in that we need to discover the overall friendship networks of students instead of predicting the existence of a single link independently. Therefore, for the aforementioned problem, a special solution framework is required.

Based on these observations, our research aims to discover an important web of students, i.e., students' class-level friendship networks. We are devoted to designing a framework for discovering friendship networks in each class through mining educational data stored in the university's management system, including ID photos (appearance feature), campus smart card records (behavior feature), course grades (intelligence feature), and psychological test scores (psychology feature). Therewith, we proposed a data-driven friendship network discovery framework, namely CANDY (**Cl**A**ss**-level frie**N**dship network **D**iscover**Y**) by taking advantage of graph learning [29]. Firstly, we represent the features of each dimension as a dense vector through representation-learning related theory. Secondly, we design a G matrix to remove the interference of redundant information in the traditional adjacency matrix on the network generation experiment. Third, we use a random shuffle strategy on data augmentation to tackle the overfitting issue caused by the small dataset. Fourth, we use a Wasserstein distance-based conditional generative adversarial network (W-CGAN) [1] as the main generative model. In other words, we aim to train a generator which can generate class-level friendship networks based on the features mentioned above. In this step, we propose a deep adversarial optimization strategy for the training of the generative adversarial networks (GAN) based model to solve the problem caused by the sparsity of the matrix. Finally, we design P_G, R_G and $F1_G$, as evaluation matrices which are the variants of precision, recall, and F1 score, to evaluate network generation comprehensively.

Our contributions could be summarized as follows:

– We design a framework for students' class-level friendship network discovery based on student information data in the education management system.

– We propose a deep adversarial optimization strategy for the GAN based model to solve the problem caused by the sparsity of adjacency matrices in the network discovery experiment.
– We conduct comprehensive experiments on the real-world dataset and the extensive results demonstrate the effectiveness of the CANDY framework.

This paper is organized as follows: In the next section, related work is reviewed briefly. The problem formulation is presented in Sect. 3. In Sect. 4, the proposed CANDY mining framework is introduced in detail. In Sect. 5, we introduce the details of our dataset. In Sect. 6, we explain the details of experiments and analyze their results. In the final section, we summarize the conclusion and future direction of our research.

2 Related Work

Research about social tie inferring attracts tremendous attention in the community of network analysis and network mining. Crandall et al. [4] propose a probabilistic framework to explore the connection between social relations and the number of co-occurrences, and demonstrated it based on the data of Flickr's users, as early as of 2010. Since then, researchers pay great attention and effort to social tie inference based on various geo-located datasets. Olteanu et al. [17] carry out an experiment to explore the effect of co-location information that stems from social relationships between users on location privacy. The above-mentioned researchers cumulatively prefer to treat social networks extracted by co-occurrence as a feature for the various applications rather than to explore the detailed relationship between co-occurrence and diverse social relationships. Deng et al. [5] simulate the whole process of online friend-making (online friendship) through the construction of game models and performed some field experiments. Liu et al. [13] and Xia et al. [28] examine citation and collaboration networks of scholars through mining publication meta-data. Liu et al. [14] propose a model based on network representation learning, namely Shifu2, to discover advisor-advisee relationships hidden behind scientific collaboration networks.

Friendships of students always gain high popularity and are considered as a special kind of social behavior. Yao et al. [33] propose a semi-supervised method to detect the friend list of students. Based on the random model, they predict the friend list according to their co-occurrence record. They verify the correctness of the friend list by using it to predict academic performance. Xu et al. [30] consider that university students' social behaviors show significant homophily in the aspect of major subject and course grade. Their research aims to find the social network of university students by eliminating the homophily effect. Khalil et al. [9] focus on the impact of course selection type on academic performance. The results demonstrate that students who took the same class with friends are more likely to achieve better grades than students who took the class alone or took the class with different students. Some studies attempt to reveal students' social relationships, like [33]. However, they barely validate the methods effectively, except for simply using students' co-presence as a characteristic to predict their achievement. Discovering student social relationships based on datasets from the education system remains an open topic.

3 Problem Formulation

In this section, we introduce notations in this paper and then formally define the research problem. Firstly, in our problem setting, we implement a directed graph $\mathcal{G} = (V, E)$ to represent the class-level friendship network, where the node in V represents students and the edge in E represents friend relationships. The edge from node A to node B represents that student A considers student B to be his or her friend. The corresponding adjacency matrix of the friendship network $\mathcal{G}$ is $\boldsymbol{A}$. For student i, we use $\boldsymbol{x}_i = [x_{i1}, x_{i2}, ...x_{in}]$ to represent his or her personal features. $\boldsymbol{X}_j$ represents the feature matrix for the class j. (The features used in this paper include appearance feature (ID photos), psychology feature (psychological test results), intelligence feature (course grades), and behavior feature (campus smart cards)). In this research, we assume that there is a mapping relationship between individual features and friendship networks, i.e., $\boldsymbol{A} = \mathcal{F}(\boldsymbol{X})$. The goal of our study is to find this function $\mathcal{F}$.

Friendship Network Discovery Problem: Assuming there is a given feature matrix $\boldsymbol{X}_j$ of a class, our aim is to discover the adjacency matrix of the corresponding friendship network $\boldsymbol{A}$.

4 Design of CANDY

In this section, we provide a specific description of the proposed framework, CANDY. This framework consists of five components: feature representation, network representation, data augmentation, generative model, and performance evaluation. The details of each component are as follows:

4.1 Feature Representation

Previous research demonstrates that the social choice of an individual student could be impacted by other students' appearance features, psychological features, intellectual features, and behavioral features [15,32,35]. In this case, all these features are used in this research for friendship network discovery. To achieve more effective information mining, we process these features through the method of representation learning (shown in Fig. 1).

Appearance Feature Representation. In this subsection, we represent students' facial features by using their ID photos. To achieve a better and effective representation, the ID photo is processed by an auto-encoder, which is a neural network model used in a wide variety of fields [16,31]. The auto-encoder is defined as follows:

$$\begin{aligned} \boldsymbol{h_{(2)}} &= f(\boldsymbol{W}_{(2)}\boldsymbol{h}_{(1)} + \boldsymbol{b}_{(2)}) \\ \boldsymbol{h_{(3)}} &= f(\boldsymbol{W}_{(3)}\boldsymbol{h}_{(2)} + \boldsymbol{b}_{(3)}) \\ &... \\ \boldsymbol{h_{(i)}} &= f(\boldsymbol{W}_{(i)}\boldsymbol{h}_{(i-1)} + \boldsymbol{b}_{(i)}), i = 1, 2, ...k \end{aligned} \tag{1}$$

where f is the activation function, and $\boldsymbol{W}_{(i)}$, $\boldsymbol{b}_{(i)}$ are the transformation matrix and the bias vector, respectively.

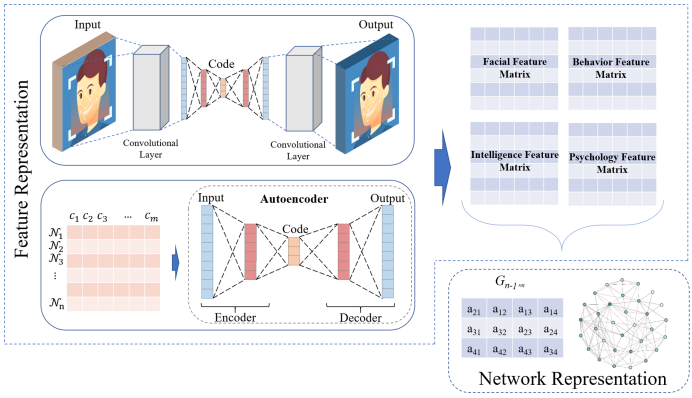

Fig. 1. Feature representation in the CANDY framework.

Behavior Feature Representation. Student behavior similarity is measured by their co-occurrence of canteen in this research [32]. Each co-occurrence is defined as the two students generating records in the canteen within a short time interval, set as 1 min [32]. The behavior feature representation is defined as follows: for student i, the similarity vector of behavior is $\boldsymbol{s}_i = [s_{i1}, s_{i2}, ...s_{ij}]$, where s_{ij} represents the co-occurrence number of student i and student j in a particular place within a specified period of time. In this case, the co-occurrence matrix $\boldsymbol{S}_k$ for class k can be defined as follows:

$$\begin{Bmatrix} 1 & s_{12} & \cdots & s_{1n} \\ s_{21} & 1 & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & 1 \end{Bmatrix}$$

Intelligence Feature Representation. In this sub-section, we use academic performance records to represent students' intelligence levels. The heterogeneity caused by the difference in the number of courses makes it difficult to be used as a feature for a machine learning model. For example, one student chooses courses A, B, and C and another student chooses courses C and D. In this case, the dimensions and contents of their academic performance features are different. To overcome the heterogeneity of students' academic performance, we employ the method mentioned in [7] for homogenization. Firstly, we embed their course through one-hot encoding and replace the 1 with the corresponding exam grade. In this way, we create the matrix $\boldsymbol{C} \in \mathbb{R}^{n \times m}$, where n and m represent the number of students and the number of courses, respectively.

However, the number of courses taken by each student is much less than the total number of courses offered by the university. For example, in our dataset, the university offers 200 courses for students, and the number of courses students take per semester is around 18. Students are taking different classes, leading to the sparsity of the matrix $\boldsymbol{C}$. To overcome this problem, we use an auto-encoder (Eq. 1) to reduce the dimension of high dimensions caused by the previous step for obtaining an effective representation of students' intelligence features.

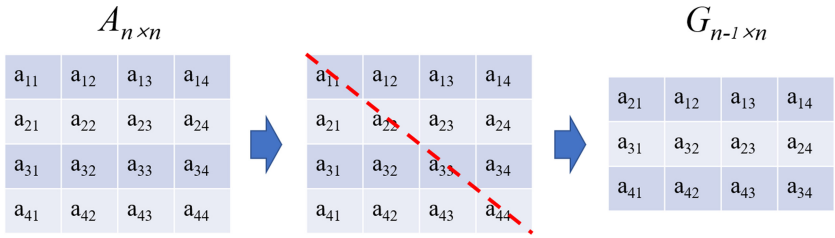

Fig. 2. G Matrix.

Psychology Feature Representation. Students' psychological characteristics are collected through the Big Five personality traits which are widely used in student-related analysis research [26,32]. The Big Five personality traits are considered as a traditional psychology model that includes five aspects: Openness, Conscientiousness, Extraversion (also often spelled as Extroversion), Agreeableness, and Neuroticism [6].

4.2 Network Representation

Generally, the adjacency matrix is used for representing the network structure. Here, we replace the adjacency matrix with its variant named G matrix, denoted by $\boldsymbol{G}$. The initial idea is that the elements on the diagonal of the adjacent matrix do not need to be learned from data as they are constant to 0. In order to eliminate the redundancy information for efficient learning, we remove the diagonal of the adjacent matrix $\boldsymbol{A}_{n\times n}$ for obtaining G matrix $\boldsymbol{G}_{n-1\times n}$. An example is shown in Fig. 2 to illustrate this process.

4.3 Data Augmentation

Unlike routine prediction experiments, each sample label in this experiment corresponds to a small-scale friendship network. In general, this kind of data is difficult to collect on a large scale. Inspired by auxiliary task design in self-supervised learning [27], we design a data augmentation method to prevent overfitting from the small dataset. The basic idea of our method is to transform the organizational form of the data without changing the rules of node connection. Suppose there are n students in a class j, we number the students and use the number as the corresponding row or column number to form the adjacency matrix $\boldsymbol{A}_{j1}$. Then we randomly shuffle the corresponding numbers of the students to regenerate the adjacency matrix $\boldsymbol{A}_{j2}$. In this case, this process is repeated n times to get n adjacency matrices $\boldsymbol{A}_{j1}$, $\boldsymbol{A}_{j2}$,...,$\boldsymbol{A}_{jn}$. (In the final experiment, these adjacency matrixes are processed by the methods in Sect. 4.2 to obtain the corresponding G matrix) For the feature matrix $\boldsymbol{X}_j$, the corresponding row would be adjusted according to the student number change for obtaining n feature matrix: $\boldsymbol{X}_{j1}$, $\boldsymbol{X}_{j2}$,...,$\boldsymbol{X}_{jn}$.

4.4 Generative Model

W-CGAN. In this research, we use a generative model to discover the topology of the friendship network i.e., this model could generate its corresponding adjacency matrix (G matrix). GAN (Generative Adversarial Networks) is a generative model based on neural network structure, which is widely used in various fields [2]. We aim to train a generator, which can generate the adjacency matrix of class-level friendship networks based on appearance features, psychology features, intellectual features, and behavioral features.

The main generative model used in this research is W-CGAN [1]. The generator in W-CGAN is the function $\mathcal{F}$ in $\mathcal{G} = \mathcal{F}(\boldsymbol{X})$ which is described in Sect. 3. Compared with classical GAN, WGAN solves the issue of vanishing gradients caused by Jensen-Shannon divergence through using smoothed Wasserstein Distance shown as follows:

$$W\left(\mathbb{P}_r, \mathbb{P}_g\right) = \inf_{\gamma \sim \Pi\left(\mathbb{P}_r, \mathbb{P}_g\right)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|] \tag{2}$$

where $\mathbb{P}_r$ is the real data distribution and $\mathbb{P}_g$ is the distribution generated by the model.

The loss function of W-CGAN is based on the Kantorovich-Rubinstein duality which is clearly described as follows:

$$\min_G \max_{D\in\mathcal{D}} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathbb{P}_r}[D(\boldsymbol{x}|\boldsymbol{y})] - \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_g}[D(\tilde{\boldsymbol{x}}|\boldsymbol{y}))] \tag{3}$$

where G and D is generator and discriminator respectively. $\mathcal{D}$ is the set of 1-Lipschitz functions. Actually, P_g is a distribution implicity defined as: $\tilde{\boldsymbol{x}} = G(z)$, where z is a combination of noises and features.

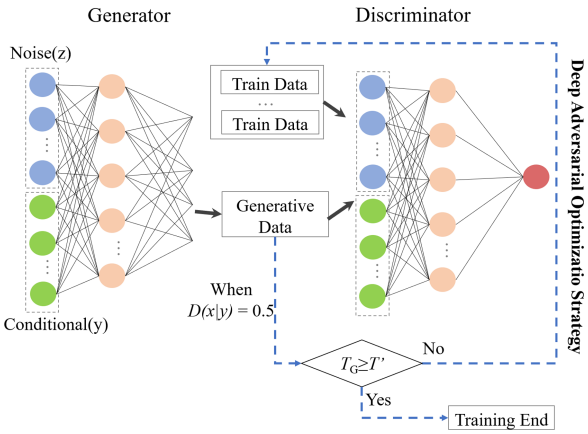

Fig. 3. The illustration of deep adversarial optimization strategy.

Deep Adversarial Optimization Strategy. Most friendship networks in the real world are sparse. Therefore, we can say there is no connection between most nodes. Likewise, it leads to most of the elements in the adjacency matrix being 0, i.e., the sparsity of the matrix. Without special treatment of sparsity, the model continuously generates all-0 matrices. In this situation, we propose an optimization strategy named as **Deep Adversarial Optimization Strategy** to overcome the target matrix's sparsity. The basic idea of this strategy is that, **lets the model learn from its own generated results to avoid the same mistakes** (shown in Fig. 3). The details are shown as follows:

1. **Step 1**: Train W-CGAN in the normal process based on real dataset R until that the discriminator is unable to differentiate between the two distributions, i.e. $D(x) = 0.5$. At this point, the generative result from the generator is $R_{G(R)}$.
2. **Step 2**: Choose an indicator T named Adversarial Indicator, which is an indicator used to evaluate the generated results from a certain aspect, and define the loop threshold value T'.
3. **Step 3**: Calculate the indicator T of $R_{G(R)}$ and define it as T_G, and then define the adversarial condition like $T_G \geq$ (or $<$) T'. If the conditions are met, the training ends. If not, put $R_{G(R)}$ into the real dataset R as a negative sample, i.e., adversarial sample, and skip back to step 1.

4.5 Performance Evaluation Methodology

Our goal is to generate the adjacency matrix corresponding to each class-level friendship network, which can be different from traditional classification and regression algorithms. Therefore, the traditional evaluation indicators like recall and precision are invalid here. To tackle this issue, we design a set of evaluation methods for network discovery: 1. We evaluate each of the generated matrices. 2. We take the mean value of the evaluation results for all generated matrices to get the final result.

First, we evaluate the individual generated matrices. Intuitively, the result of the subtraction of the two matrices should be used as the evaluation criterion for the generative model. However, the sparseness of the data makes such evaluation methods ineffective. In this case, we introduce a confusion matrix to evaluate a single matrix sample. For the adjacent matrix, matrix elements only include zero and one. According to the labels of real data and their corresponding predicted results, we divide matrix elements into four categories: TO (True One), FO (False One), TZ (True Zero), and FZ (False Zero) and define the corresponding confusion matrix. Based on this confusion matrix, we define evaluation metrics for each generated matrix: P_G (G precision), R_G (G recall), and $F1_G$ (G F1-score) as follows:

$$P_G = \frac{TO}{TO + FO} \tag{4}$$

$$R_G = \frac{TO}{TO + FZ} \tag{5}$$

$$F1_G = (1 + \beta^2)\frac{P_G \cdot R_G}{(\beta^2 \cdot P_G) + R_G} \tag{6}$$

In this work, $F1_G$ is the harmonic mean of R_G and P_G ($\beta = 1$). Then, we take the mean value of the evaluation results for all generated matrices to get the final results.

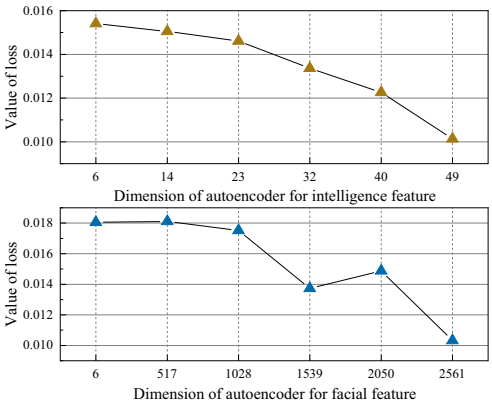

Fig. 4. The results of feature representation for facial feature and intelligence feature.

5 Dataset

The dataset used in this research includes 512 university students from the same Chinese university and all of them are freshmen who just finished their first semester exams. All participants are required to be more than 18-year-old freshmen (aged 18–20, mean = 19.03, SD = 0.21), who live in several specific freshman residential buildings (next to each other) in the same area. They come from 16 different classes. We use the data stored from the education management system, including their appearance feature (ID photos), psychology feature (psychological test results), intelligence feature (course grades), and behavior feature (campus smart card records) for discovering class-level friendship networks. This research was approved by the university's ethics committee.

The details of our dataset are introduced as follows:

1. Friendship Network Data. We use a questionnaire-based method for our data collection method in which each student is asked to write about 5–8 good friends. We collect friend relationships for each class in turn. Students from the same class are called to the lab and complete a friend relationship questionnaire. This activity takes place at the beginning of the student's second semester. Monetary compensation is given for participation in the study.
2. ID Photo Data. Our experiment utilizes personal one-inch photographs submitted by all participants at the time of enrollment in the university.
3. Academic Performance Data. Students' academic performance is generally recorded as the exam grade of each course. The academic performance data used in this research includes 13,234 records.
4. Campus Smart Card Data. In most universities, the smart card is used as a recognition tool for identifying a student. Generally, smart cards can be used for any scene of the behavior of university students and thus record tons of data for student behavior, such as bathing and eating. Financial data includes 259,513 records.

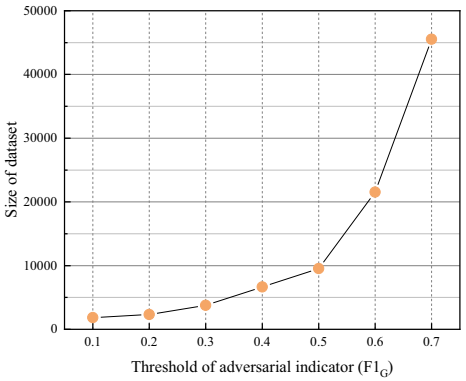

Fig. 5. The relation between the threshold of adversarial indicator and the size of the dataset.

5. Psychology Test Data. All university students are required to take psychological tests during their enrollment. We use that data to profile the psychological characteristics of the students involved in this experiment, which includes 30,720 records.

6 Experiments

6.1 Experimental Settings

For our research experiments, we use the friendship networks of twelve classes to train our model and the other four classes for testing. For the training set, we extend one network to thirty networks by using the data augmentation method mentioned in Sect. 4.3. In this case, we have 360 network samples in the training set in total. As discussed above, auto-encoders are used to obtain effective feature representation of facial features and intelligence features. We then test the performance on different dimensions (shown in Fig. 4) and the value of the loss function of the auto-encoder fluctuates slightly. Thus, we choose 6 as the final embedding dimension for computational efficiency.

In addition, as far as we know, there is no experiment exactly the same as our experiment. In this case, to verify the effectiveness of our framework, we design a comparison experiment based on binary classification: We consider the generation process of a friendship network with n nodes as n^2 independent binary classification experiments. We make predictions about whether every two students are friends or not. For example, if a class has 32 students, we make 1024 (32×32) binary classification experiments to build an adjacency matrix of their friendship network (992 binary classifications for G Matrix). The classifiers used in this part are shown as follows:

- SVM (Support Vector Machine) [24]: SVM is a classic classification algorithm and is widely used in the field of data mining.
- XGBoost [3]: XGBoost is a boosting-tree-based method and is widely used in various data mining scenarios with good performance.

– DNN (Deep Neural Networks): DNN is a trendy model based on a multi-layer neural network and is widely used in various scenarios. (Our research implements a common three-layer neural network model)

The training set and testing set used for these binary experiments are synchronous with the main model proposed in this research.

6.2 Analysis of Results

The results of our generation experiments are shown in this part. The adversarial indicator T of the CANDY framework used in this paper is $F1_G$ with adversarial condition $F1_G \geq 0.7$ (The reasons are given below). The results of the CANDY framework and all comparison experiments are shown in Table 1. First, the CANDY framework proposed in this paper outperforms all comparison algorithms. The possible reason is that friendship between people is not independent of others. For example, A is a friend of B, and A is also a friend of C. Therefore, B and C may also be friends because they have mutual friends. In this case, independent binary classification experiments fail to catch these high-order network characteristics.

Table 1. Performance of all comparison experiments. CANDY ($T_{(F1_G)} = 0.4$) represents the experiment results based on the CANDY framework when the adversarial indicator T is $F1_G$ and the threshold is set to 0.4.

	P_G	R_G	$F1_G$
SVM	0.26976	0.26412	0.26691
XGBOOST	0.30108	0.29758	0.29932
DNN	0.42192	0.43217	0.42693
CANDY ($T_{(F1_G)} = 0.4$)	0.33499	0.33721	0.33609
CANDY ($T_{(F1_G)} = 0.7$)	0.52717	0.50165	0.51409

Although experiments verify the effectiveness of the CANDY framework, the overall performance is still not good enough as expected but the best F1-score is only 0.51409. The main reasons behind this are as follows.

First, each sample in this experiment corresponds to a matrix, so the complexity of the experiment is greatly increased. Second, this experiment is equivalent to conducting 992 interconnected binary classification tasks at the same time. For such a challenging task, the dataset used in this research is too small to support good performance. Although we propose a data augmentation method to mitigate the impact of the small-scale dataset, such a linear method can not completely solve this problem. More data involved in model training will help to achieve better results. Thirdly, as the purpose of this paper is to emphasize the effectiveness of the CANDY framework instead of pursuing high accuracy, we did not collect enough features for the experiment. For example, we only use the co-occurrence of the cafeteria as a feature to infer the friendship network. Co-occurrence in more places can also effectively help in improving the

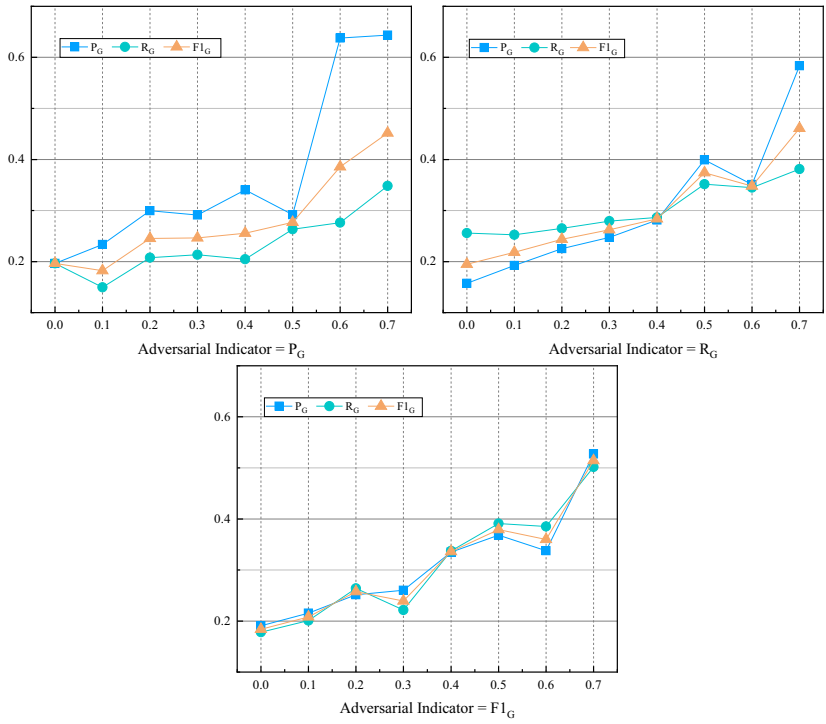

Fig. 6. The performances of CANDY with different adversarial conditions.

experimental results, e.g. that research by [30] uses the co-occurrence of 17 places to infer the friendship network. A more comprehensive feature is bound to result in higher performance.

In addition, the threshold of indicator T is set at 0.7 because the size of the training set increases with the time of training. According to the deep adversarial optimization strategy proposed in this research paper, 'half-finished products' from the generator will consistently be added to the training set as long as the requirement is not met, leading to that concerning the training set will increase with the time of training. For example, if we choose $F1_G$ as the adversarial indicator, the changing trend of the size of the training set is shown in Fig. 5. An extensive training set leads to longer model training times. In this case, the adversarial indicator T is set to 0.7 to balance the training process and the algorithm's performance.

Moreover, the selection of adversarial indicator T is crucial for our algorithm and its efficiency because the indicator can significantly impact the algorithm's performance. We perform experiments to analyze this issue. We set the adversarial indicators as P_G, R_G, and $F1_G$, separately, and for each adversarial indicator, we take values from 0.1 to 0.7 to test its impact on performance. We use P_G, R_G, and $F1_G$ to quantify the results, and the results are shown in Fig. 6. It can be observed that different indicators have different effects. The higher the threshold set on the training set, the better the performance

on the test set. The experimental results demonstrated that $F1_G$ is an optimal candidate for the adversarial indicator, which is consistent with the intuition that $F1_G$ is affected by both P_G and R_G.

Table 2. Performance with different variants. For concise presentation in the table, we use shorthand to represent each part of CANDY framework: R (Raw data), A (Data augmentation), W-CG (W-CGAN), $\boldsymbol{G}$ (G matrix).

	P_G	R_G	$F1_G$
R+W-CG	0.08912	0.12418	0.10377
R+A+W-CG	0.18057	0.17394	0.17719
R+A+W-CG+$\boldsymbol{G}$	0.19056	0.17794	0.18403
CANDY ($T_{(F1_G)} = 0.4$)	0.33499	0.33721	0.33609
CANDY ($T_{(F1_G)} = 0.7$)	0.52717	0.50165	0.51409

Besides, we compare the performance of the CANDY framework with its variants as well, and the results are shown in Table 2. Note that the W-CGAN means ordinary W-CGAN without a deep adversarial optimization strategy. First, feeding the raw data into W-CGAN achieves an inferior performance, because the raw dataset only contains 16 networks, causing serious overfitting issue. By contrast, the improvement of 'R+A+W-CG' demonstrates the effectiveness of our data augmentation. The G matrix improves the performance but not sharply. Nevertheless, it contributes to the training efficiency of the model by reducing redundant information. Finally, we added the deep adversarial optimization strategy (i.e., CANDY framework) with $T = F1_G$ and the performance is greatly improved. In other words, the experimental results in this part demonstrate the effectiveness of each part of the CANDY framework.

7 Conclusion

In this work, we propose a data-driven framework to discover students' class-level friendship networks based on student data from the education management system, including students' ID photos, psychological test results, course grades, and the record of campus smart cards. First, we represent features as low-dimensional dense vectors through representation learning. Secondly, we use conditional GAN with Wasserstein distance as the main generative model and propose a deep adversarial optimization strategy to tackle the problem caused by the sparsity of adjacency matrices. Finally, we transplant the evaluation system of classification experiments into the network generation for achieving a comprehensive evaluation. The performance on a real-world dataset validates the effectiveness of the proposed framework. In future work, in order to achieve a comprehensive understanding of the social patterns among students, a series of indicators will be explored to quantify students' social patterns from a dynamic and static perspective, respectively.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the International Conference on Machine Learning 2017, pp. 214–223. PMLR (2017)
2. Chen, J., Li, Y., Ma, K., Zheng, Y.: Generative adversarial networks for video-to-video domain adaptation. In: Proceedings of the 32th AAAI Conference on Artificial Intelligence, pp. 3462–3469. AAAI Press (2020)
3. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM (2016)
4. Crandall, D.J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., Kleinberg, J.: Inferring social ties from geographic coincidences. Proc. Natl. Acad. Sci. **107**(52), 22436–22441 (2010)
5. Deng, X., Song, D., Wei, L.: A dynamic game model analysis for friendship selection. J. Intell. Fuzzy Syst. **1**, 1–9 (2019)
6. Goldberg, L.R.: An alternative 'description of personality': the big-five factor structure. J. Pers. Soc. Psychol. **59**(6), 1216 (1990)
7. Guo, T., et al.: Graduate employment prediction with bias. In: Proceedings of the 32th AAAI Conference on Artificial Intelligence, pp. 670–677. AAAI Press (2020)
8. Hernández, I., Rivero, C.R., Ruiz, D.: Deep web crawling: a survey. World Wide Web **22**(4), 1577–1610 (2019)
9. Khalil, L.J., Khair, M.G.: Social network analysis: friendship inferred by chosen courses, commuting time and student performance at university. Int. J. Reason.-based Intell. Syst. **10**(1), 59–67 (2018)
10. Lande, D., Fu, M., Guo, W., Balagura, I., Gorbov, I., Yang, H.: Link prediction of scientific collaboration networks based on information retrieval. World Wide Web **23**(4), 2239–2257 (2020)
11. Liu, J., et al.: Artificial intelligence in the 21st century. IEEE Access **6**, 34403–34421 (2018)
12. Liu, J., et al.: Data mining and information retrieval in the 21st century: a bibliographic review. Comput. Sci. Rev. **34**, 100193 (2019)
13. Liu, J., Ren, J., Zheng, W., Chi, L., Lee, I., Xia, F.: Web of scholars: a scholar knowledge graph. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2153–2156 (2020)
14. Liu, J., et al.: Shifu2: a network representation learning based model for advisor-advisee relationship mining. IEEE Trans. Knowl. Data Eng. **33**(4), 1763–1777 (2021)
15. Morelli, S.A., Ong, D.C., Makati, R., Jackson, M.O., Zaki, J.: Empathy and well-being correlate with centrality in different social networks. Proc. Natl. Acad. Sci. **114**(37), 201702155 (2017)
16. Muhammed, Fatih, B., Abubakar, A., James Y, Z.: Concrete autoencoders for differentiable feature selection and reconstruction. In: Proceedings of the 36th International Conference on Machine Learning, pp. 444–453. PMLR (2019)
17. Olteanu, A.M., Huguenin, K., Shokri, R., Humbert, M., Hubaux, J.P.: Quantifying interdependent privacy risks with location data. IEEE Trans. Mobile Comput. **16**(3), 829–842 (2016)
18. Overgoor, J., Adamic, L.A., et al.: The structure of us college networks on Facebook. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, pp. 499–510 (2020)
19. Parkinson, C., Kleinbaum, A.M., Wheatley, T.: Similar neural responses predict friendship. Nat. Commun. **9**(1), 332 (2018)

20. Pickering, T.A., et al.: Diffusion of a peer-led suicide preventive intervention through school-based student peer and adult networks. Front. Psychiatry **9**, 598 (2018)
21. Ream, G.L.: The interpersonal-psychological theory of suicide in college student suicide screening. Suicide Life-Threat. Behav. **46**(2), 239–247 (2016)
22. Ren, J., et al.: Matching algorithms: fundamentals, applications and challenges. IEEE Trans. Emerging Top. Comput. Intell. **5**(3), 332–350 (2021)
23. Rodríguez-Triana, M.J., Prieto, L.P., Holzer, A., Gillet, D.: Instruction, student engagement, and learning outcomes: a case study using anonymous social media in a face-to-face classroom. IEEE Trans. Learn. Technol. **13**(4), 718–733 (2020)
24. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2001)
25. Van Duijn, M.A., Zeggelink, E.P., Huisman, M., Stokman, F.N., Wasseur, F.W.: Freshmen into a friendship network. J. Math. Sociol. **27**(2–3), 153–191 (2003)
26. Vedel, A.: The big five and tertiary academic performance: a systematic review and meta-analysis. Pers. Individ. Differ. **71**, 66–76 (2014)
27. Wang, X., He, K., Gupta, A.: Transitive invariance for self-supervised visual representation learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1329–1338 (2017)
28. Xia, F., Liu, J., Ren, J., Wang, W., Kong, X.: Turing number: how far are you to AM Turing award? ACM SIGWEB Newsl. (Autumn), 1–8 (2020)
29. Xia, F., et al.: Graph learning: a survey. IEEE Trans. Artif. Intell. **2**, 109–127 (2021)
30. Xu, J.Y., Liu, T., Yang, L.T., Davison, M.L., Liu, S.Y.: Finding college student social networks by mining the records of student id transactions. Symmetry **11**(3), 307 (2019)
31. Xu, W., Tan, Y.: Semisupervised text classification by variational autoencoder. IEEE Trans. Neural Netw. Learn. Syst. **31**(1), 1–14 (2019)
32. Yao, H., Lian, D., Cao, Y., Wu, Y., Zhou, T.: Predicting academic performance for college students: a campus behavior perspective. ACM Trans. Intell. Syst. Technol. (TIST) **10**(3), 24 (2019)
33. Yao, H., Nie, M., Su, H., Xia, H., Lian, D.: Predicting academic performance via semi-supervised learning with constructed campus social network. In: Candan, S., Chen, L., Pedersen, T.B., Chang, L., Hua, W. (eds.) DASFAA 2017. LNCS, vol. 10178, pp. 597–609. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-55699-4_37
34. Yin, H., Yang, S., Song, X., Liu, W., Li, J.: Deep fusion of multimodal features for social media retweet time prediction. World Wide Web **24**, 1027–1044 (2020)
35. Zhang, D., et al.: Judging a book by its cover: the effect of facial perception on centrality in social networks. In: Proceedings of the Web Conference 2019, pp. 2290–2300. ACM (2019)
36. Zhou, Y., et al.: Extracting representative user subset of social networks towards user characteristics and topological features. World Wide Web **23**(5), 2903–2931 (2020)