Indexed Metrics for Link Prediction in Graph Analytics

Marcus Lim[1], Azween Abdullah[2], NZ Jhanjhi[3]
School of Computer Science & Engineering (SCE)
Taylor's University
Selangor, Malaysia

Mahadevan Supramaniam[4]
Research and Innovation Management Centre
SEGI University, Malaysia

Abstract—**With the explosive growth of the Internet and the desire to harness the value of the information it contains, the prediction of possible links (relationships) between key players in social networks based on graph-theory principles has garnered great attention in recent years. Consequently, many fields of scientific research have converged in the development of graph analysis techniques to examine the structure of social networks with a very large number of users. However, the relationship between persons within the social network may not be evident when the data-capture process is incomplete or a relationship may have not yet developed between participants who will establish some form of actual interaction in the future. As such, the link-prediction metrics for certain social networks such as criminal networks, which tend to have highly inaccurate data records, may need to incorporate additional circumstantial factors (metadata) to improve their predictive accuracy. One of the key difficulties in link-prediction methods is extracting the structural attributes necessary for the classification of links. In this research, we analysed a few key structural attributes of a network-oriented dataset based on proposed social network analysis (SNA) metrics for the development of link-prediction models. By combining structural features and metadata, the objective of this research was to develop a prediction model that leverages the deep reinforcement learning (DRL) classification technique to predict links/edges even on relatively small-scale datasets, which can constrain the ability to train supervised machine-learning models that have adequate predictive accuracy.**

***Keywords*—Link prediction; social network analysis; criminal network; deep reinforcement learning**

I. Introduction

The rapid accessibility of the internet and social media platforms has resulted in the exponential growth of social networks. Thus, providing a medium for the gathering of internet users with common interests would increase the number of possible associations and facilitate the establishment of new communities.

The discovery of new links is a valuable attribute of the friend recommender systems employed by social network platforms. A number of algorithms employed by IBM within its own internal social network were investigated by Chen et al. [1], who found them to facilitate the establishment of connections among its employees. The technique of predicting the presence of hidden/missing links or the formation of new structural connections is usually described as a link-prediction problem in social networks.

In social network analysis (SNA), the problem of predicting links poses an ongoing challenge and represents a key research topic [2]. There are two main approaches for predicting links between nodes within a network. One approach is based on the features of the nodes and the other is based on the topological properties of the connected nodes within the network. In the context of social network platforms, users are represented as nodes or vertices and user-related information or profile attributes may not be easily accessible. The choice to use topological properties for link prediction is preferable as it is mainly based on models derived from graph-theory analysis. Even though the topological properties of real-world social networks may not always be consistent, classical link prediction metrics based on topological properties, for example, common neighbours [3] and Adamic/Adar [4], seldom factor metadata as weights when formulating SNA metrics.

The ability to predict links accurately has many valuable applications in a range of domains that can be modelled using a network-oriented structure. In the field of bio-informatics, link prediction is used to identify the structure of connecting proteins [5]. Link prediction is also applied in e-commerce to develop recommendation systems [6]. In the domain of criminal-network analysis (CNA), link prediction is critical for the swift identification of key terrorist or criminal groups [7]. As the problem of predicting links is pertinent to a wide range of domains, many algorithms have been explored in recent years to address this issue. Many of these algorithms have relied mainly on classical machine-learning techniques that require training on the relevant features of a large dataset to achieve adequate predictive accuracy.

In SNA, the topological properties of network-oriented domains are considered along with environmental factors that can have an impact on changes in the links or relationships among users over time [8]. These environmental factors, commonly referred to as metadata, such as judicial convictions, arrest records and community crime rates in the context of criminal networks, furnish supporting information that may subsequently be used to shape the structural configurations of the networks [9].

This research received the funding under Taylor's Flagship Project under project # (TUFR/2017/004/06) from the Centre for Research Management, Taylor's University, Malaysia.

Deep reinforcement learning (DRL) is a machine-learning (ML) algorithm that utilises the reinforcement-learning (RL) model and incorporates a deep-learning (DL) algorithm to serve as a function approximator. In the latest developments, a DRL model has been successfully shown to be capable of self-learning across multiple domains by assimilating layers of feature learning via self-simulation based only on the provision of basic rules applicable to a particular domain [10, 11].

A DL algorithm that formulates an ML model processes multiple layers of feature learning that are typically progressively extracted from a sufficiently large dataset to acquire a precise abstraction of the domain properties. DL, which is also referred to as a deep neural network (DNN), functions as an artificial neural network (ANN) by imitating the learning process of brain neurons [12-14]. DL automates representation learning by the abstraction of domain features via layers of an ANN from the input to the output layer. The use of DNN in the formulation of ML models therefore minimises its reliance on human input to programme feature abstraction rules for specific domains.

RL is an ML framework that involves the development of programmes that function as agents that use trial and error to learn to navigate within an environment to achieve pre-defined goals [15]. These agents are guided towards achieving a goal by performing a series of tasks, which when completed are assessed as either a success or failure by a system of domain-related rules. Successes are usually given positive marks as rewards and negative marks as punishments.

Our research is expected to contribute to the development of a new set of enhanced indexed link-prediction metrics that can better predict missing links than classical link-prediction metrics. An evaluation experiment was performed on a time-series criminal-network dataset. This model was developed by indexing SNA metrics with metadata to enhance the capability of law enforcement agencies to more accurately identify critical unknown relationships in criminal networks. The model proposed in this research may have some limitations in that it is constructed based on relatively small dataset which are characteristics of criminal or terrorist networks compared to social networks such as Facebook. The relatively small dataset may have an impact on the predictive performance of certain machine learning models being trained.

In the rest of this paper, we our research work is presented as follows: In Section II, we review relevant research work involving ML models that incorporate weighted metrics. In Section III, we describe our development of the proposed and baseline models and the training methodology used. In Section IV, we describe the properties of the dataset and the experimental setup and discuss the experimental results. We present our research conclusions in Section V and in Section VI we consider the trajectory of subsequent research work.

II. Related Work

Supervised ML algorithms are usually the preferred techniques for solving link-prediction problems. ML was first reported in 2003 by Liben-Nowell and Kleinberg [16] based on their research on the value of the structural attributes of graphs and their development of models trained on bibliographic datasets. In 2006, Hasan et al. conducted research [7] based on the technique developed by Liben-Nowell and Kleinberg. Subsequently, many other researchers have developed models using the same technique. A majority of the models proposed by these researchers were trained and evaluated on co-authored or bibliographic datasets [16], [17], [18]. Song et al. developed link-prediction models trained on a feature matrix based on node similarity and proximity measures extracted from large-scale real-world datasets such as MySpace and Facebook for use in matrix factorisation [19].

In 2011, Zaki and Al Hasan conducted a survey and provided a review of other link-prediction approaches based on linear algebra, Bayesian probabilistic models and Bayesian networks [20].

Cukierski, et al. [21] constructed a model for predicting links based on the random-forest classification technique based on an extraction of 94 graph features. The results from their model trained on Flickr datasets were found to achieve a high level of predictive accuracy.

In the development of a highly accurate link-prediction algorithm, it is critical to compute set feature metrics derived from the structural properties of graph datasets. As the accuracy of such models also depends on the use of large-scale datasets, the generation of a feature matrix requires significant computer resources. Social networks such as Facebook, which had some 700 million users in 2011 and a monthly average incremental of 20 million, poses a considerable challenge in terms of computer resources [22]. Furthermore, the structural configurations of these networks also exhibit certain attributes, for example, a power-law degree distribution [23] and small-world properties [24]. These properties must be considered when local structural features are being computed as nodes, which requires huge computer resources. As such, a subgraph is used.

Silver et al. made a major contribution to the advancement of DRL research with their development of AlphaGo, a programme that plays the ancient strategic board game Go. AlphaGo, which incorporates the Monte Carlo tree search (MCTS) technique, achieves accurate super-intuitive judgement by identifying the scope of analyses that have the highest likelihoods of success [25]. The game of Go, which is considered by artificial intelligence experts to be the holy grail in the field of computer science, has more possible variations in board positions than all the atoms in the universe that are visible to man. In 2017, AlphaGo demonstrated its extreme intelligence by beating the world’s best professional Go player by a clear margin.

Silver et al. achieved another milestone in artificial general intelligence with the development of AlphaGo Zero. With only the basic rules of the games provided, AlphaGo Zero succeeded in mastering different two-player games with complete information such as Go, Shogi and Chess, using its self-learning algorithm to compete with different versions of itself [26].

Using self-simulated dataset generated within three days for training purposes, AlphaGo Zero defeated AlphaGo with a score of 3–0.

Unlike classical supervised ML techniques, such as the random forest, DRL has the capability of being trained on a self-generated dataset via self-play. As a result, the DRL technique is relevant to the modelling of network-oriented domains with datasets that are comparatively small, for example criminal syndicates, using a self-simulated dataset.

The research journals we reviewed offer scant evidence of any examination of ML models for predicting links that integrate SNA metrics indexed with metadata measurements and the DRL technique with respect to dynamic criminal networks. In this research, we conducted experiments to fill this research gap in the construction of models trained on a time-series criminal-network dataset.

III. Models and Methodology

In this research, we constructed two DRL models for link prediction: a baseline model for predicting links using just classical SNA metrics (BSNA-DRL) and another that uses classical SNA metrics indexed with metadata weights (ISNA-DRL).

A. *Baseline Model SNA (BSNA-DRL) using Classical Link-Prediction Metrics*

1) Common Neighbours (CN)

$$CN_{xy} = |\Gamma(x) \cap \Gamma(y)| \qquad (1)$$

The CN score of the node pair x and y, *CN$_{xy}$*, denotes the number of directly connected nodes that are common to x and y. $\Gamma(x)$ and $\Gamma(y)$ denote the set of directly connected nodes of both x and y [27].

2) Jaccard Coefficient (JC)

$$JC_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \qquad (2)$$

The JC score of the node pair x and y, *JC$_{xy}$*, denotes the number of neighbours common to x and y as a ratio of the total number of directly connected nodes of both x and y [27].

3) Adamic/Adar measure (AA)

$$AA_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z} \qquad (3)$$

The AA score of the node pair x and y, *AA$_{xy}$*, denotes the summation of the inverse value of the degree *k* of node *z*, which is the neighbour common to all directly connected nodes of both x and y [27],[28].

4) Preferential Attachment (PA)

$$PA_{xy} = \Gamma_x \text{ x } \Gamma_y \qquad (4)$$

The PA score of the node pair x and y, *PA$_{xy}$*, indicates the probability that two nodes will be connected, which is proportional to the degree of the nodes [27].

B. *Indexed Model SNA (ISNA-DRL) using Classical Link-Prediction Metrics Factored with Metadata Weights*

In this research, classical link-prediction metrics were factored with two types of metadata to measure their impact on the precision of the link-prediction model compared with that of the baseline model. The metadata used included the number of criminal records and education level. Persons associated with criminals with lengthy criminal records are expected to form a relationship in the future, and those with a low education level who are associated with criminal networks are expected to have a higher likelihood of forming a relationship in the future.

1) Indexed Common Neighbours (iCN)

$$iCN_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\log(\mathrm{md1}_z) + \log(\mathrm{md2}_z)}{2} \qquad (5)$$

The indexed CN score of the node pair x and y, *iCN$_{xy}$*, denotes the common neighbour node *z* that is factored by the weighted average of the metadata, i.e., the md1 and md2 value attributes of node *z*. The *iCN$_{xy}$*, value increases with the likelihood of a link forming between nodes x and y.

2) Indexed Jaccard Coefficient (iJC)

$$iJC_{xy} = \frac{\displaystyle\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\log(\mathrm{md1}_z) + \log(\mathrm{md2}_z)}{2}}{|\Gamma(x) \cup \Gamma(y)|} \qquad (6)$$

The indexed JC score of the node pair x and y, *iJC$_{xy}$*, denotes the *iCN$_{xy}$* value as a ratio of the total number of directly connected nodes of both x and y. The *iJC$_{xy}$*, value increases with the likelihood of a link forming between nodes x and y.

3) Indexed Adamic/Adar measure (iAA)

$$iAA_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\log(\mathrm{md1}_z) + \log(md2_z)}{2}$$

$$\text{x} \frac{1}{\log k_z} \qquad (7)$$

The indexed AA score of the node pair x and y, i*AA$_{xy}$*, denotes the summation of the weighted average of the metadata (md1, md2) value attributes of the common neighbour node *z*, factored with the inverse value of the degree *k* of node *z*.

4) Indexed Preferential Attachment (iPA)

$$iPA_{xy} = \sum_{\mathrm{x}' \in \Gamma_x} \frac{\log(\mathrm{md1}_{x'}) + \log(\mathrm{md2}_{x'})}{2}$$

$$\text{x} \sum_{\mathrm{y}' \in \Gamma_y} \frac{\log(\mathrm{md1}_{y'}) + \log(\mathrm{md2}_{y'})}{2} \qquad (8)$$

The indexed PA score of the node pair x and y, i*PA$_{xy}$*, indicates that the probability that two nodes will be connected is proportional to the degree of nodes x and y after factoring the weighted average of the metadata (md1, md2) value attributes of the neighbouring nodes x' and y' of node x and y, respectively.

C. Proposed ISNA-DRL Model

The proposed metadata-indexed SNA link-prediction CNA model (ISNA-DRL) (Fig. 1), which represents an extension of the research on link prediction in the criminal network domain, applies the MCTS method in network searches [29-31]. To assess the predictive precision of the ISNA-DRL model, we used the area under the curve (AUC) score [32].

The ISNA-DRL model leverages a value network (indexed-SNA-metrics neural net) (Fig. 1), which is a DNN trained with features extracted from indexed SNA metrics. During the training of the model, the indexed SNA feature matrix extracted from the number of criminal records and education levels were factored as a score for the indexed-SNA-metrics neural net. The indexed-SNA-metrics neural net is a function approximator that generates output values used to rank each pair of nodes based on the likelihood of links forming or disappearing. The MCTS commences its tree search from the pair of nodes with the maximum combined indexed-SNA-metrics score estimated by the value network. The cumulative scores obtained by the RL agent from all the completed simulated network instances are then fed back to the neural net for re-calibrating the ISNA-DRL model's hyper-parameters to improve its predictive accuracy in subsequent iterations (Fig. 1).

Notes (Fig. 1):

a) The topological features of the criminal-network dataset are used to compute the indexed SNA values.

b) Features are extracted from the metadata for computation of the indexed SNA metrics.

c) The indexed SNA metrics of the criminal-network dataset are used to formulate the features matrix.

d) The indexed SNA feature matrix derived from the metadata features, e.g., the number of arrest records and the education levels, are processed by the indexed-SNA-metrics value network.

e) The indexed-SNA-metrics value network approximates the node pairs that will most likely to form a link or have their link disappear.

f) The MCTS commences to traverse the network by initiating a network-instance simulation by identifying links/edges, which are then ranked in accordance with their probability scores (P_0, P_1) as approximated by the indexed-SNA-metrics value network.

g) The simulation of the network states, S_0 to S_N, occurs by the roll-out of the link-prediction process by the MCTS policy network. These simulated network states are assessed regarding their prediction accuracy by comparing their results with the test datasets (T_1 to T_{10}).

h) The predictive accuracy scores of the predicted instances of the network are processed by the RL agent and fed back to the indexed-SNA-metrics neural net to adjust the related hyper-parameters and improve the accuracy of subsequent simulations.

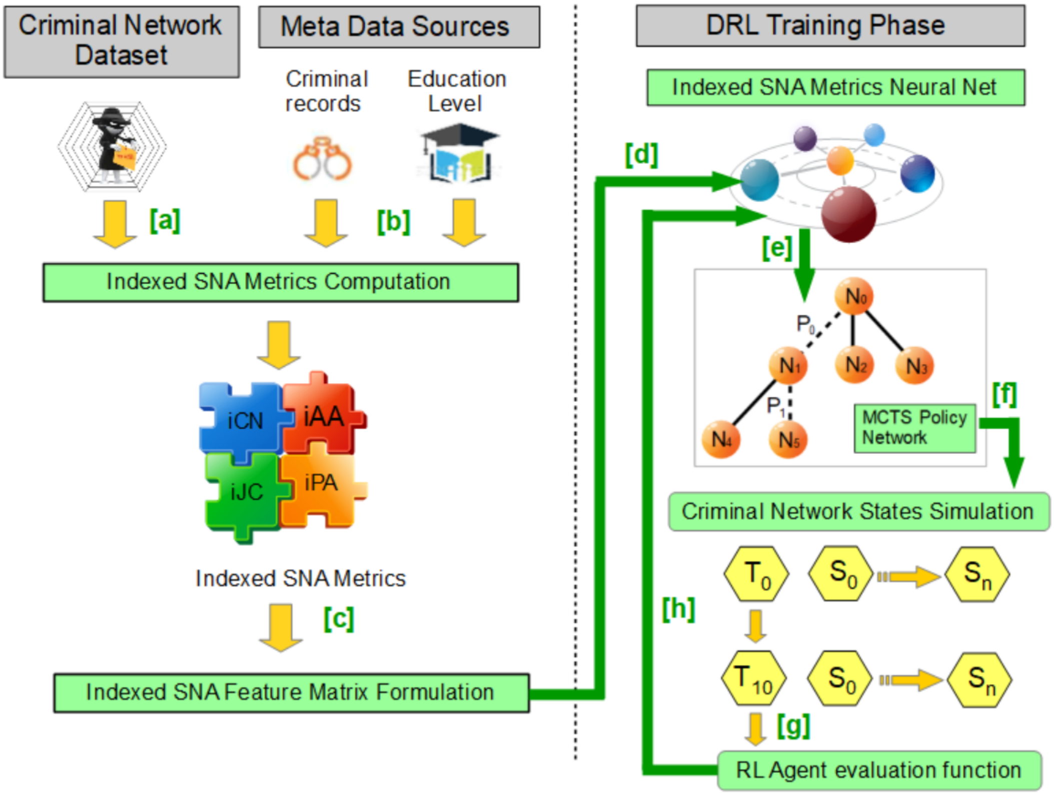

Fig 1. Proposed ISNA-DRL Link-Prediction Model based on SNA and Metadata Metrics.

The SNA measurements selected for the formulation of the indexed SNA feature matrix included the CN, AA, JC and PA metrics [33]. The indexed SNA metrics derived (*iCN_{xy}*, *iJC_{xy}*, *iAA_{xy}*, *iPA_{xy}*) (5-8) from the structural properties of the network and metadata were extracted for representation learning by the value network to approximate the ranking score of the links. These scores were then processed by the MCTS function to generate the most probable network instances. The indexed SNA scores are formulated as feature vectors, which are stored as data records. This array of features is then used to train the indexed SNA value network to perform a binary classification of the links/edges predicted to have either positive or negative labels. The link/edge that is predicted to be most likely to form is given a positive label and that predicted to be most likely to disappear is given a negative label.

The training of the ISNA-DRL model involves the representation learning of arrays of features, which comprise scores of indexed SNA metrics that denote the probability of the formation or disappearance of links/edges in the future. The cumulated value of the indexed SNA prediction metrics are calculated for every link in each feature array. The ISNA-DRL model was evaluated using the test dataset to ascertain whether the trained model predicted network instances with the required accuracy (Fig. 1).

The MCTS network tree search process creates a network instance for each link that is predicted by the policy network to be most likely to change in the following iteration. Each iteration of the MCTS network traversal, which starts from the root node, creates an initial simulated instance, S_0, from which the search process continues to the next node, navigating in accordance with the score of a positive or negative link most likely to form.

The generation of a probable network instance, S_2, as a result of a new link being predicted from the current state, S_1, is the outcome of the navigation of the RL agent to S_2 from S_1, based on the rules of the default policy network. The prediction made regarding a new network instance based on the likely formation of a new link is determined by the values of the SNA prediction metrics used to rank the links. When a simulation has been completed, every simulated network state is assessed against the original network dataset to determine the accuracy of the prediction. Any variances found from this assessment are evaluated by a cost function to adjust the hyper-parameters of the indexed-SNA-metrics neural net and the MCTS function to achieve a better prediction. These hyper-parameters are then incorporated into the subsequent network-instance simulation in accordance with the link-prediction rules (Fig. 1).

The link prediction accuracies of the BSNA-DRL and ISNA-DRL models constructed on classical SNA metrics (*CN_{xy}*, *JC_{xy}*, *AA_{xy}*, *PA_{xy}*) (1-4) and indexed SNA metrics (*iCN_{xy}*, *iJC_{xy}*, *iAA_{xy}*, *iPA_{xy}*) (5-8), respectively, were evaluated based on their AUC indices. The AUC index of an ML model indicates the precision of the modelling process in identifying the underlying domain patterns, with the score ranging from 0 to 1. The higher is the AUC index achieved by the model, the more accurate are its predictions likely to be.

D. Metadata Indexing

Metadata indexing refers to the process of factoring the measurements of various pieces information obtained from the environment into SNA measurements for link prediction, which can influence the precision of the links predicted by the models. With reference to the criminal-networks domain, metadata may include the number of criminal records, education level and age, which can shape the structural configurations of a dynamic network and affect the underlying metrics on which link predictions are based [34]. In the proposed ISNA-DRL model (Fig. 1), the number of criminal records and the education levels of the members of the criminal network were factored into the feature matrix formulated from the indexed SNA metrics (5-8), which were then used to train the indexed SNA value network. The output of the indexed SNA value network is an approximation of a set of ranked scores that identify the node pairs with the highest likelihood of changing over time.

E. Time-Series Dataset

The graph algorithm is used in modelling network-oriented domains that evolve over time, such as online social or criminal groups, whose topological configurations may vary with time [34]. Participants in a network, which are denoted as nodes, may enter or exit the group as time passes. The structural configurations of the network may also change, for example, when the strength of the relationships or links among the participants change over time. The dynamic nature of such real-world networks are reflected in a time-series dataset.

IV. Experiments and Results

In this research we used a time-series dataset of the Caviar drug import syndicate [35]. This dataset contains a series of eleven time-series snapshots of arrest raids conducted to seize drugs from the criminal network over a 2-year period. We evaluated both the proposed ISNA-DRL and baseline BSNA-DRL models based on their AUC indexes, as these values are not skewed by the presence of imbalanced classes and this method is typically employed to evaluate the accuracy of ML classifier models.

A. Experiment Setup

The BSNA-DRL and ISNA-DRL prediction models were trained using a multidimensional feature matrix which was computed based on classical SNA metrics (*CN_{xy}*, *JC_{xy}*, *AA_{xy}*, *PA_{xy}*) (1-4) and indexed SNA metrics (*iCN_{xy}*, *iJC_{xy}*, *iAA_{xy}*, *iPA_{xy}*) (5-8), respectively, by factoring the metadata features derived from the Caviar dataset. This computation combines classical SNA link-prediction metrics with metadata indices to derive values that represent the probability of the formation of positive or negative links at each of the time-series snapshots of the Caviar dataset (Fig. 1).

We randomly divided the time-series dataset into training and test datasets with an 80:20 proportion, respectively. Of the eleven time-series snapshots of arrest raids, ten were used as the training set from which a random selection of positive edges was made. Then, a random selection of negative edges was made until the numbers of negative and positive edges were equal. Each of the snapshots from the time-series test dataset was then processed by the model to predict the likely network topology from the test dataset.

The tenth and eleventh time-series snapshot of the Caviar dataset (Fig. 2 and 3) depict the actual evolution of the network during the interval. The eleventh time-series snapshot was used for testing to determine the precision of the BSNA-DRL and ISNA-DRL models in making predictions.

The experiment was conducted in two stages. In the first stage, we used the feature matrix computed for each classical SNA measurement to train both the models. The objective of this step was to determine the impact of factoring metadata on the prediction properties of each SNA measurement made independently from the others. Fig. 4 shows the predictive accuracies of the BSNA-DRL and ISNA-DRL models based on the use of individual SNA metrics.

In the next stage of the experiment, we combined all four classical metrics to formulate a feature matrix, which was used to train both models.

B. Results and Discussion

The ISNA-DRL model was able to identify more links/edges (Fig. 6) than the BSNA-DRL model (Fig. 5), and these links/edges were expected to change in the eleventh time-series snapshot when compared with the actual time-series snapshot at T_{11} (Fig. 2) and T_{10} (Fig. 3).

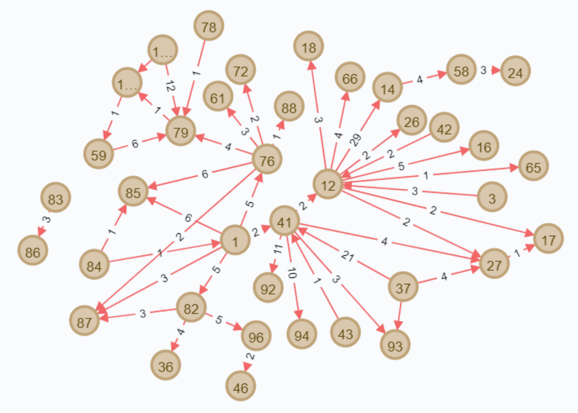

Fig 2. Actual Criminal Network at Time-Stamp T_{11}.

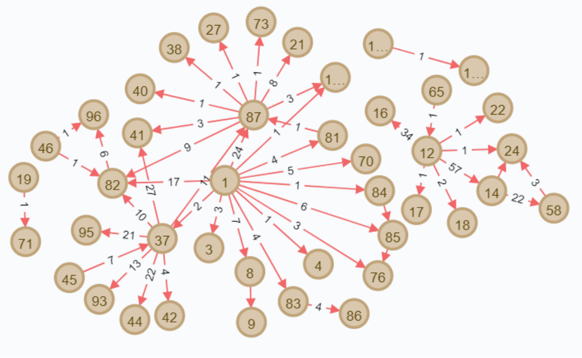

Fig 3. Actual Criminal Network at Time-Stamp T_{10}.

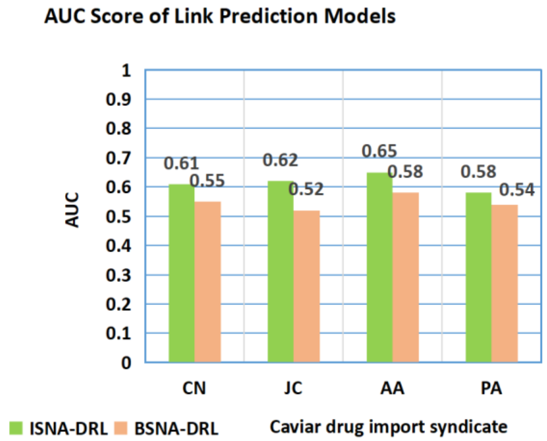

Fig 4. AUC Scores of Models Built with Individual SNA Metrics.

Although the predictions of the ISNA-DRL model for four links/edges were incorrect, i.e., node pairs (76,87), (1,81), (14,24), (12,14) (Fig. 6), the BSNA-DRL model incorrectly predicted five more links/edges, i.e., node pairs (41,93), (37,82), (41,87), (1,83) and (46,82) (Fig. 5).

A comparison of the predicted structural configurations of the Caviar network at the eleventh time step, T_{11}, with the experimental results indicates that the indexed SNA metrics (iCN_{xy}, iJC_{xy}, iAA_{xy}, iPA_{xy}) (5-8) of the ISNA-DRL model (Fig. 6) achieved better prediction accuracy than the BSNA-DRL model (Fig. 5).

The AUC scores of the ISNA-DRL model that uses individual classical SNA metrics such as CN, JC AA and PA (Fig. 4) to factor metadata scores are better by 0.06, 0.10, 0.07 and 0.04, respectively, than the AUC scores obtained by the BSNA-DRL model (Fig. 4), which did not incorporate metadata indexing. This indicates that metadata indexing does not cause inconsistent results when all the indexed SNA metrics are combined.

The better performance of the ISNA-DRL model than the BSNA-DRL model seemed to be related to the use of the SNA metrics indexed with metadata (iCN_{xy}, iJC_{xy}, iAA_{xy}, iPA_{xy}) (5-8), which reflect the real-life characteristics of criminal activity.

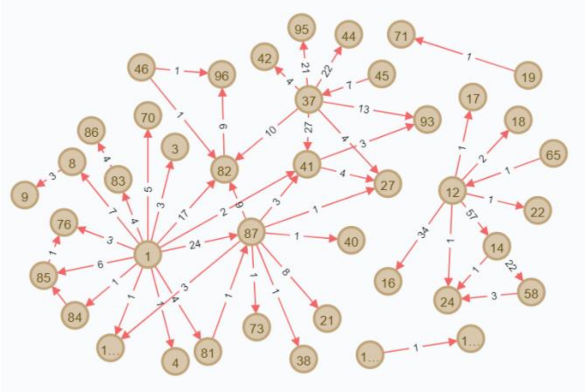

Fig 5. Network Predicted by BSNA-DRL Model.

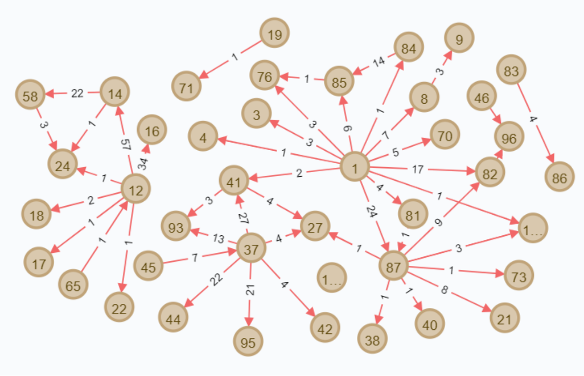

Fig 6. Predicted Network by ISNA-DRL Model.

To determine the overall prediction precisions of the models, we also conducted an experiment in which we combined all four classical SNA metrics to train both the proposed ISNA-DRL and BSNA-DRL models (Fig. 7 and 8).

The results indicate that the use of metadata scores also improved the precision of the predictions made by the ISNA-DRL trained with indexed SNA feature matrix. The overall improvement of the ISNA-DRL model using the combined indexed SNA metrics (*iCN*$_{xy}$, *iJC*$_{xy}$, *iAA*$_{xy}$, *iPA*$_{xy}$) (5-8) could be due to the fact that metadata scores provide further information related to the nodes, which influences the selection of node pairs towards those with a higher likelihood of forming positive or negative links in the future.

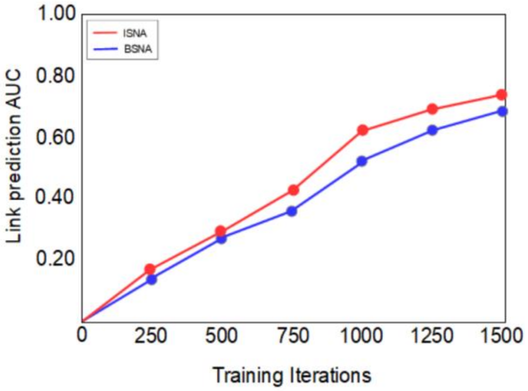

Fig 7. AUC Metrics of Link-Prediction Models with Combined SNA Metrics.

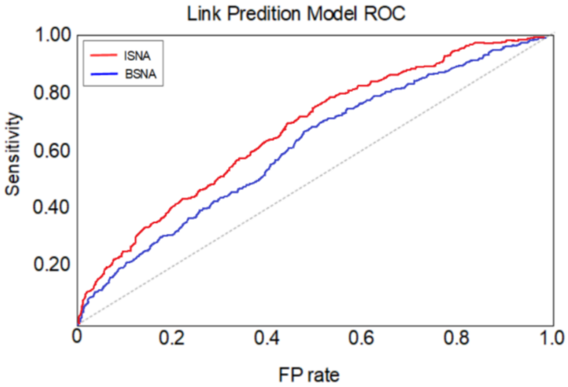

Fig 8. ROC Curve of Link-Prediction Model with Combined SNA Metrics.

In general, the experiment conducted to evaluate the link-prediction performance associated with metadata indexing indicated that the ISNA-DRL model, which factors metadata scores with SNA link-prediction metrics, performed better than the baseline BSNA-DRL model (Fig. 7 and 8), which was trained with a feature matrix formulated without metadata indexing and simulated self-generated datasets using the DRL technique.

The experimental results for the proposed DRL link-prediction model, which was trained on a time-series criminal-network dataset, are consistent with those obtained by Lim, Marcus et al. [30], [31].

In both [30] and [31] (Table II), the experiments were also conducted with link prediction models constructed using relatively small time-series dataset and leveraging on DRL.

In [30], the link prediction model, TDRL-CNA, only use features formulated from classical SNA metrics to train the model. However, in the TDRL-CNA model, additional SNA features metrics, i.e. Hub Index and Preferential Attachment index were formulated as weights in the hidden layers of the SNA metrics neural network. This model which uses breath first search (BFS) ranking algorithm, performed with approximately with the same level of predictive accuracy (AUC score of 0.78) compared to our ISNA-DRL model (AUC score of 0.74) (Table II). This result seem to indicate that the ISNA-DRL model despite being a more simplified model compared to the TDRL-CNA model, was able to achieve a comparable level of performance by using SNA metrics which were indexed by metadata.

TABLE I. AUC SCORES OF BSNA-DRL LINK-PREDICTION MODEL AND ISNA-DRL MODELS

Model	AUC	Time-score(Hr)	Iterations
BSNA-DRL	0.69	1.12	1500
ISNA-DRL	0.74	1.85	1500

TABLE II. COMPARISON OF DRL LINK PREDICTION MODELS FROM RELATED RESEARCH WORKS

Model	ISNA-DRL	TDRL-CNA	MDRL-CNA
ML technique	DRL	DRL	DRL with metadata fusion
Tree search ranking algorithm	MCTS	BFS	MCTS
SNA metrics	metadata indexed	classical	classical
Dataset	11 time-periods	11 time-periods	20 time-periods
Maximum nodes	42	27	55
Training time-score (hour)	1.85	Not available	4.3
Training iterations	1500	1500	2500
AUC Score	0.74	0.78	0.79
Authors	Current work	[30]	[31]

In [31], the link prediction model, MDRL-CNA, incorporated meta data features in the formulation of the feature matrix used to train the neural networks of the model instead of factoring into the SNA metrics computation. This model was able to achieve a performance which is better than the ISNA-DRL model with an AUC score of 0.79. However, due to additional complexity of fusing metadata in the weight formulation by leveraging on DL, additional computing resources of 4.3 hours were used to train the MDRL-CNA model compared to the 1.85 hours required to train the ISNA-DRL model. Therefore, considering the resources required to train the models, the factoring of meta data into the formulation of SNA metrics may prove a viable option in constructing a link prediction model where time is a constraint.

V. Conclusion

In the experiments conducted in this study, the link-prediction model constructed from combined indexed SNA metrics that factored metadata scores (ISNA-DRL model) performed consistently better than the BSNA-DRL model that did not factor metadata scores. This result is supported by the respective AUC scores of 0.74 and 0.69 achieved by the ISNA-DRL and BSNA-DRL models (Table I). The experimental results also indicate that models constructed by leveraging the DRL technique can be successfully trained on smaller and self-generated datasets.

The incorporation of metadata, i.e., criminal records and education level, with classical SNA metrics enhanced the predictive precision of the ISNA-DRL algorithm, which is likely due to the incorporation into the model of real-life factors that may shape criminal-network behaviour. The improved predictive accuracy of the proposed model can contribute significantly to disrupting the activities of criminal syndicates.

VI. Future Work

Future research should focus on the investigation of the results obtained when more than two metadata scores are factored with SNA metrics in the construction of a link-prediction algorithm, which is expected to increase the precision of the ISNA-DRL model. However, the factoring of SNA metrics with an increased number of metadata scores must be explored regarding its ability to either further improve the accuracy of the ISNA-DRL model or diminish its predictive performance due to over-fitting.

Reference

[1] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, “Make new friends, but keep the old: recommending people on social networking sites,” in Proceedings of the 27th international conference on Human factors in computing systems, ser. CHI ’09. New York, NY, USA: ACM, 2009, pp. 201–210. [Online]. Available: http://doi.acm.org/10.1145/1518701.1518735

[2] Wang, Peng, BaoWen Xu, YuRong Wu, and XiaoYu Zhou. "Link prediction in social networks: the state-of-the-art." Science China Information Sciences 58, no. 1: 1-38, 2015.

[3] Aich, Suman, and Anita Mehta. "The clash of the Titans: How preferential attachment helps the survival of the smallest." The European Physical Journal Special Topics 223, no. 13: 2745-2758, 2014.

[4] Güneş, İsmail, Şule Gündüz-Öğüdücü, and Zehra Çataltepe. "Link prediction using time series of neighborhood-based node similarity scores." Data Mining and Knowledge Discovery 30, no. 1: 147-180, 2016.

[5] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, “Mixed membership stochastic block models for relational data with application to protein-protein interactions,” Proceedings of International Biometric Society-ENAR Annual Meetings, 2006.

[6] Ricci, Francesco, Lior Rokach, and Bracha Shapira. "Recommender systems: introduction and challenges." In Recommender systems handbook, pp. 1-34. Springer, Boston, MA, 2015.

[7] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki, “Link prediction using supervised learning,” SDM Workshop of Link Analysis, Counterterrorism and Security, 2006.

[8] Hajibagheri, Alireza, Gita Sukthankar, and Kiran Lakkaraju. "Leveraging network dynamics for improved link prediction." In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, pp. 142-151. Springer, Cham, 2016.

[9] Arnoux, Thibaud, Lionel Tabourier, and Matthieu Latapy. "Combining structural and dynamic information to predict activity in link streams." In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 935-942. 2017.

[10] Li, H., Kumar, N., Chen, R., & Georgiou, P. “A Deep Reinforcement Learning Framework for Identifying Funny Scenes in Movies”. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 3116-3120, 2018.

[11] Bahdanau, Dzmitry, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. "End-to-end attention-based large vocabulary speech recognition." In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4945-4949. IEEE, 2016.

[12] Zeiler, M.D., & Fergus, R. “Visualizing and Understanding Convolutional Networks”. ECCV, Vol.1, 2014.

[13] Chen, Xi-liang, Lei Cao, Chen-xi Li, Zhi-xiong Xu, and Jun Lai. "Ensemble network architecture for deep reinforcement learning." Mathematical Problems in Engineering 2018 (2018).

[14] Duan, Y., Chen, X., Houthooft, R., Schulman, J., & Abbeel, P. “Benchmarking Deep Reinforcement Learning for Continuous Control”. ICML.Vol.1.2016.

[15] Shi, Yangyang et al. "Contextual spoken language understanding using recurrent neural networks." In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5271-5275. IEEE, 2015.

[16] D. Liben-Nowell and J. Kleinber, “The link-prediction problem for social networks,” Journal of the American Society for Information Science and Technology, 58, no. 7, 2007.

[17] Anitha, M. V., and Linda Sara Mathew. "Dynamic Link Prediction in Biomedical Domain." In International Conference On Computational Vision and Bio Inspired Computing, pp. 219-225. Springer, Cham, 2019.

[18] Gupta, Anshul, Shalki Sharma, and Hirdesh Shivhare. "Supervised Link Prediction Using Forecasting Models on Weighted Online Social Network." In Proceedings of International Conference on ICT for Sustainable Development, pp. 249-261. Springer, Singapore, 2016.

[19] H. H. Song, T. W. Cho, V. Dave, Y. Zhang, and L. Qiu. “Scalable proximity estimation and link prediction in online social networks,” in Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference, ser. IMC ’09. New York, NY, USA: ACM, 2009, pp.322–335. [Online].Available: http://doi.acm.org/10.1145/ 1644893. 1644932.

[20] M. A. Hasan and M. J. Zaki, Social Network Data Analytics, C. C. Aggarwal, Ed. Springer, 2011.

[21] W. J. Cukierski, B. Hamner, and B. Yang, “Graph-based features for supervised link prediction,” International Joint Conference on Neural Networks, 2011.

[22] Insidefacebook”http://www.insidefacebook.com/2011/06/12/ facebooksees-big-traffic-drops-in-us-and-canada-as-it-nears-700-million-usersworldwide/.

[23] Stai, Eleni, Vasileios Karyotis, and Symeon Papavassiliou. "A hyperbolic space analytics framework for big network data and their applications." IEEE Network 30, no. 1: 11-17, 2016.

[24] Grabow, Carsten, Stefan Grosskinsky, Jürgen Kurths, and Marc Timme. "Collective relaxation dynamics of small-world networks." Physical Review E 91, no. 5: 052815, 2015.

[25] D.Silver et al. “Mastering the Game of Go without Human Knowledge”. Nature, 550, no. 7676: 354–359, 2017.

[26] D. Silver et al., “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play”, Science, 362, no. 6419, 1140-1144, DOI. 10.1126. science. aar6404, 2018.

[27] Mallek, Sabrine, Imen Boukhris, Zied Elouedi, and Eric Lefèvre. "Evidential link prediction in social networks based on structural and social information." Journal of computational science 30: 98-107, 2019.

[28] L.A. Adamic, E. Adar, Friends and neighbors on the web, Social Networks 25, 211–230, 2003.

[29] Marcus Lim, Azween Abdullah, NZ Jhanjhi. “Performance Optimization of Criminal Network Hidden Link Prediction Model with Deep Reinforcement Learning", Journal of King Saud University - Computer and Information Sciences, 2019.

[30] Lim, Marcus, Azween Abdullah, Nz Jhanjhi, Muhammad Khurram Khan, and Mahadevan Supramaniam. "Link Prediction in Time-Evolving Criminal Network With Deep Reinforcement Learning Technique." IEEE Access 7 (2019): 184797-184807.

[31] Lim, Marcus, Azween Abdullah, N. Z. Jhanjhi, and Muhammad Khurram Khan. "Situation-Aware Deep Reinforcement Learning Link Prediction Model for Evolving Criminal Networks." IEEE Access, 2019.

[32] Lim, Marcus, Azween Abdullah, NZ Jhanjhi, and Mahadevan Supramaniam. "Hidden Link Prediction in Criminal Networks Using the Deep Reinforcement Learning Technique." Computers 8, no. 1: 8, 2019.

[33] Özcan, Alper, and Şule Gündüz Öğüdücü. "Multivariate temporal link prediction in evolving social networks." In 2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), pp. 185-190. IEEE, 2015.

[34] Holme, Petter. "Modern temporal network theory: a colloquium." The European Physical Journal B 88, no. 9: 234, 2015.

[35] Borgatti, S.P., Everett, M.G. and Freeman, L.C. 2002. Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies. 1, 2002.