Link Prediction using Social Network Analysis over Heterogeneous Terrorist Network

Akash Anil, Durgesh Kumar, Shubhanshu Sharma, Rakesh Singha,
Ranjan Sarmah, Nitesh Bhattacharya and Sanasam Ranbir Singh
Department of Computer Science and Engineering
Indian Institute of Technology Guwahati, Assam, India 781039
Email: (a.anil, k.durgesh, s.shubhanshu, r.singha, ranjan23, nitesh2012, ranbir) @iitg.ernet.in

***Abstract*—Social network analysis (SNA) has been effectively used in counter-terrorism analysis by generating homogeneous network. In this paper, we consider a large dataset reporting various terrorist attacks over the globe and represent the dataset as a heterogeneous network. The objective of this paper is to the explore the effect of various link prediction frameworks such as topic modeling, network topology and graph kernels. We propose bipartite based link prediction over topic feature relationship, heterogeneous version of node proximity based link prediction and graph kernel methods. From various experimental observation, it is evident that bipartite method based on topic modeling also return comparable results (sometimes better) as that of node proximity and graph kernel.**

***Keywords*-Social Network Analysis, Heterogeneous, Terrorist Networks, Graph Kernels, Topic Modeling.**

I. Introduction

With the increase in availability of digitized data related to terrorist activities, social network analysis has garnered increasing attention in analyzing counter-terrorism related data in recent time. Several works are presented in literature to understand terrorist network using various social network analysis methods such as link prediction [1], structural analysis [2]–[4], modeling [5], [6] etc. Major concerns in all these studies are (i) the size of the datasets and (ii) the nature of the datasets. Majority of the datasets used in prior studies are small and the underlying networks are homogeneous in nature (nodes are of same type i.e., nodes of the network are mostly either terrorists or terrorist organizations). Getting a large terrorist network of high quality is one of the core challenges in academic research because of several reasons as described in [2]; *size, incompleteness, fuzzy boundary* and *dynamics*. Recently, various agencies have put up efforts to create large databases related to terrorist activities in public domain. Some of such datasets are the *Global Terrorist Data* (GTD) [1], *South Asian Terrorist Portal* (SATP) [2] etc. However, these databases provide the information in semi-structured or unstructured forms. Constructing homogeneous network out of such dataset is a non-trivial task. Considering the nature of such datasets, it is important to explore the methods which can deal with the semi-structure or unstructured nature of the datasets.

An event of a terrorist attack is often defined by heterogeneous set of attributes such as terrorist organizations involved, accused terrorists, place of attack, target type, material used, victims etc. Analysis related to counter-terrorism often needs to study relationship between different attributes such as *terrorist group and target type*, or *material used and target type* or *country and potential terrorist organizations* etc. Since relationship between such attributes are often influenced by various other attributes, it is important to take other attributes into consideration while exploring hidden relationship between different attributes. Motivated by the above two factors - *unstructured form of data and heterogeneous set of attributes*, the objective of this paper is to predict links (hidden or future) between different attributes of the event of terrorist attacks by considering others attributes into account and study the underlying network structures. This paper studies three approaches to incorporate multi-attributes and study their effects; heterogeneous network, spectral transformation and topic modeling on link prediction. First, we consider heterogeneous network, because it enables us to exploit heterogeneous attributes and at the same time allows us to apply existing proximity based link prediction methods such as *Common Neighbour (CN), Jaccard Coefficient (JC) [7], Adamic-Adar Index (AA) [8], Resource Allocation (RA) [9] etc.* Proximity based methods consider only the explicit relationships present in the network. In the second approach, we explore latent relationships between attributes by applying spectral graph kernels such as *Path Counting kernel [10], Neumann kernel [11] and Exponential kernel [12].* Though the above two approaches attempt to explore both the explicit and latent relationships, they will be greatly affected by underlying representation i.e., graphical representation. Majority of the publicly available datasets are either in semi-structured or unstructured forms. Considering these constraints, it may be important to study methods which will be able to process unstructured dataset. This has motivated us to investigate the effect of Latent Dirichlet Allocation *(LDA)*; a generative based topic modeling method which can effectively process unstructured text dataset.

[1] http://www.start.umd.edu/gtd/
[2] http://www.satp.org/

978-1-5090-1893-2/15 $31.00 © 2015 IEEE
DOI 10.1109/SmartCity.2015.82

In this study, we investigate the effect of the above three frameworks over GTD dataset; a large semi structure dataset reporting different terrorist attacks over 4 decades. Our analysis is mainly focusses on; (i) link prediction problem between heterogeneous attributes, and (ii) verifying causal effect of other attributes. From various experimental analysis, it is evident that all three frameworks return promising performances over heterogeneous dataset. To the best of our knowledge, this is the first study that explores heterogeneous attributes through heterogeneous network, graph kernel and topic modeling over GTD dataset. This paper therefore reports performance of above three methods only.

The rest of the paper is organized as follows. Section II presents a brief literature survey and background studies. Section III describes the proposed frameworks. about the used methods and experimental setups. It initializes with briefing the details about dataset used, its preprocessing, which is followed by describing particular methods and evaluation. Results and Discussion are presented in Section IV, which is followed by conclusion and exploring the future works in Section V.

II. Background and Related Studies

Social network analysis over terrorist network came into popularity after the attack of September 11, 2001. However, Sparrow [2] explored the application of social network analysis in criminal intelligence in 1991. In this paper, Sparrow examined six different notions of *centrality* and three major notions of *equivalence* for their relevance in revealing the mechanics and vulnerabilities of criminal enterprises. In 2001, John Arquilla and David Ronfeldt [13] analyzed the terrorist and criminal organizations using graph theoretic approach. Valdis Krebs [3] utilized the centrality measures and different SNA tools to understand the structure of a terrorist network generated using news publications. This work predicted the important terrorist and their future allies for 9/11 attack. In a similar way, Josep A. Rodrguez [14] prepared and analyzed a terrorist network for the attack of March 11 on Spain. Marc Sageman [15] used SNA on AlQaeda and found that terrorists were divided mainly into four clusters on based on regions. All the studies discussed above modeled a homogeneous terrorist network. However, real world terrorist networks are heterogeneous in nature. Considering this fact, Kathleen Carley [16] proposed Dynamic Network Analysis (DNA), which used multi-agent modeling to destabilize terrorist network. In DNA, the underlying graph is multi-modal and multi-relational, which is capable of capturing more information as compared to homogeneous network.

The $n(x)$ is the set of neighbour of x and the $n_p(x)$ is the set of neighbour nodes with priority p of node x.

Table I
Heterogeneous Transformation of Node Proximity based Similarity Measures.

	Homogeneous	**Heterogeneous**
CN	$S_{x,y} = \lvert n(x) \cap n(y) \rvert$	$S_{x,y} = \sum_{p \in P} p.\lvert n_p(x) \cap n_p(y) \rvert$
JC [7]	$S_{x,y} = \frac{\lvert n(x) \cap n(y) \rvert}{\lvert n(x) \cup n(y) \rvert}$	$S_{x,y} = \sum_{p \in P} p.\frac{\lvert n_p(x) \cap n_p(y) \rvert}{\lvert n_p(x) \cup n_p(y) \rvert}$
AA [8]	$S_{x,y} = \sum_{z \in n(x) \cap n(y)} \frac{1}{log(\lvert n(z) \rvert)}$	$S_{x,y} = \sum_{p \in P} \sum_{z \in n_p(x) \cap n_p(y)} p.\frac{1}{log(\lvert n_p(z) \rvert)}$
RA [9]	$S_{x,y} = \sum_{z \in n(x) \cap n(y)} \frac{1}{\lvert n(z) \rvert}$	$S_{x,y} = \sum_{p \in P} \sum_{z \in n_p(x) \cap n_p(y)} p.\frac{1}{\lvert n_p(z) \rvert}$

A. Heterogeneous Networks

In this paper, a heterogeneous network is defined by a four tuple $G =< V, E, P, \omega >$, where V is the set of vertices, E is the set of edges, P is the set of possible node weights and ω is a function $\omega : V \rightarrow P$. If $|P| = 1$, then $\omega(x) = \omega(y), \forall x, y \in V$. We refer this kind of graph as a homogeneous network and it is realized by a graph $G =< V, E >$.

B. Heterogeneous Measure of Similarity

Traditional node proximity based similarity measures such as CN, JC, AA, and RA are defined over homogeneous networks. In this paper, we propose heterogeneous counterparts of these measures. Table I summarizes the proposed heterogeneous similarity measures. It enables us to assign different priorities to different types of nodes/attributes. For example, the common neighbour between nodes x and y is defined as $\sum_{p \in P} p|n_p(x) \cap n_p(y)|$ where $n_p(x)$ denotes the set of neighbour nodes having priority p of node x. Higher the number of common neighbours with higher priority, higher is the similarity score.

We further propose a heterogeneous interpretation for the graph kernels. To incorporate the priority weights of participating nodes, we consider harmonic mean of the node pairs and modify the adjacency matrix as follow. Let $\mathbf{A}$ be the adjacency matrix representation for the given network, then $\mathbf{M}$ (modified adjacency matrix), can be derived as

$$\mathbf{M_{x,y}} = hm(\omega(x), \omega(y))\mathbf{A_{x,y}}$$

where $hm(\omega(x), \omega(y))$ is the harmonic mean of importance/weights of the nodes x and y i.e., $\frac{2.\omega(x).\omega(y)}{\omega(x)+\omega(y)}$. Table II summarizes the spectral transformation [17] for heterogeneous graph kernel.

C. Topic Modeling

Topic modeling [18] is a statistical approach to explore hidden themes in large document collection. Latent Dirichlet Allocation *(LDA)* forms the basis for large number of topic modeling framework. It has been applied for extracting hidden pattern in news corpus [19], scientific documents [20]

Table II
Different spectral kernels on graph and their spectral transformation, K(A): Kernel on matrix A

Kernel	Kernel Function(Homogeneous)	Kernel Function(Heterogeneous)	Spectral Kernel Transformation (Heterogeneous)
Path Counting [10]	$K(\mathbf{A}) = \mathbf{A}^k = \mathbf{U}\Lambda^k\mathbf{U}^T$	$K(\mathbf{M}) = \mathbf{M}^k = \mathbf{U}\Lambda^k\mathbf{U}^T$	$K(\mathbf{M}) = \mathbf{U}F(\Lambda^k)\mathbf{U}^T$ where $F(\Lambda^k)$ is defined by $f(\lambda) = \alpha\lambda$ and $\alpha > 0$
Exponential kernel [12]	$K(\mathbf{A}) = \exp(\mathbf{A}) = \sum_{i=0}^{\infty} \frac{\alpha^i}{i!}\mathbf{A}^i = e^{\alpha\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathbf{T}}$	$K(\mathbf{M}) = \exp(\mathbf{M}) = \sum_{i=0}^{\infty} \frac{\alpha^i}{i!}\mathbf{M}^i = e^{\alpha\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathbf{T}}$	$K(e^{\alpha\mathbf{M}}) = \mathbf{U}F(e^{\alpha\mathbf{\Lambda}})\mathbf{U}^T$ where $F(e^{\alpha\mathbf{\Lambda}})$ is defined by $f(\lambda) = e^{\alpha\lambda}$
Neumann kernel [11]	$K(\mathbf{A}) = (\mathbf{I} - \alpha\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathbf{T})^{-1}$ where $\alpha^{-1} > \lambda_1$	$K(\mathbf{M}) = (\mathbf{I} - \alpha\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathbf{T})^{-1}$ where $\alpha^{-1} > \lambda_1$	$K(\mathbf{M}) = [\mathbf{U}F(\mathbf{I} - \alpha\mathbf{\Lambda})\mathbf{U}^T)]^{-1}$where $f(\lambda) = \frac{1}{1-\alpha\lambda}$ defines the kernel.

[21], etc. The intuition behind LDA is that each document $(d_1, d_2, ...d_n)$ is created based on hidden theme called *topics* $(t_1, t_2, ...t_K)$. LDA models each document as a random mixture of latent topics, where each topic is characterized as probability distribution(multinomial) over all words in the vocabulary set [18]. A Dirichlet prior α is assumed to model the topic-word multinomial distribution for each topic and similarly another Dirichlet prior β is assumed to model the document-topic multinomial distribution for each document.

1) **Generative Process of LDA:** Figure 1 depicts the plate representation of LDA Model. The generative process of LDA is described in Algorithm 1. For a given document, it first chooses a multinomial distribution θ_d of topics in it , using Dirichlet prior α. Given the multinomial distribution of topics (θ_d) in the document, it first chooses a topic and then chooses a word according to multinomial distribution of word in the given topic.

Algorithm 1: Generative Process of LDA [18]

```
Choose N ~ Poisson(ξ)
Choose $\theta_d$ ~ Dir($\alpha$)
for all N words in the document $d$ do
    Choose a topic $z_n$ ~ Multinomial($\theta_d$)
    Choose a word $w_n$ from $p(w_n|z_n,\beta)$, a multinomial
    probability conditioned on the topic $z_n$
end for
```

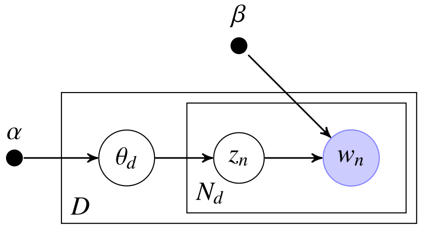

Figure 1. Plate notation of LDA. [18]

2) **Topic modeling in Social Network Analysis:** Topic modeling methods has been used for solving various Social Network problems like Community Detection, etc . Several variants of LDA like Author Recipient Topic models *(ART)* and Role Author-Recipient-Topic *(RART)* [22] has been used for determining roles and topics in enron and academic emails. Community Role Author Topic *CART* model [23] given by Pathak et al. for community detection in Social Network exploits both link structure and contents. Gaston et al. [24] use Topic modeling and HITS algorithm to find community based key extraction from Dark Web portal.

III. Methods and Experimental Setup

Experiments have been carried out by taking equal weights to all the attributes for all the link prediction methods. For graph kernels, we also see the effects of different weights for different attributes. It is because, with other methods (proximity based and topic modeling) the data preprocessing becomes much complex.

A. Dataset

The dataset *Global Terrorist Data (GTD)* used in this work is collected from *National Consortium for the Study of Terrorism and Responses to Terrorism*, University of Maryland [25]. This repository contains information about more than 140,000 terrorist attacks across the globe collected over a span of year 1970 to 2014. The first column of the dataset represents the occurrence of particular terrorist attack as event-id, having information about day, month and year. Other important informations such as victim country(country), modus-operandi of terrorist groups(attack type, target type, weapon type), name of the terrorist groups etc, are represented by other columns.
For this work, we extract the following ten important columns as features; country, region, provstate, city, attacktype1, targtype1, targsubtype1, gname, weaptype1, and weapsubtype1. We generate a heterogeneous graph by drawing an edge between these features, if they are associated with the same event/attack. Table III, gives the exact number of nodes and edges considered for experiments. All the entries corresponding to terrorist group's name as "unknown/Unknown", are deleted from the dataset.

B. Data preparation for Node proximity based similarity measures

In this paper, all the four local-similarity measures namely, Common Neighbour, Jaccard Coefficient, Adamic Adar index, Resource Allocation discussed in Section II have been explored for link prediction. The above heterogeneous graph is divided into two parts, i.e. Training and Testing. Training graph is prepared accumulating data between the years 1970 to 2009. Data between the year 2010 to 2014 is considered as test graph. In this work, we assume that no new nodes have been introduced during the period 2010 to 2014.

Table III
Dataset (Global Terrorist Data, 1970-2014)

Method	Node (Training)	Node (Validation)	Node (Test)	Edge (Training)	Edge (Validation)	Edge (Test)
Node Proximity	17326	NA	1783	2197702	NA	38499
Graph Kernel	17326	17326	1783	1961173	2197702	38499
Topic Modelling	17326	NA	1783	2197702	NA	38499

Method: The adjacency matrix representation is supplied as input to the above discussed local link prediction algorithms. For all the models, we get an output matrix, which gives the similarity score between node pairs. Since, we give equal importance to all the attributes, so the weight ω, behaves as just the scaling factor on output similarity matrix, having no effect on relative importance.

C. Data Preparation for Graph Kernels

A global similarity measure corresponds to learning useful information from the whole network. Most of the methods defined in literature are based on the finding sum of number of paths of different length between the node pairs. Using an adjacency matrix($\mathbf{A}$), for graph $\mathbf{G}$, it is clear that, these models evaluate a power series. Graph kernels discussed above, can be transformed in the power series and can be used as a link prediction model based on global similarity basis.
We divide the given graph into three parts with their adjacency representation, namely, Training $\mathbf{A_s}$(from 1970-2004), Validation $\mathbf{A_v}$(1970-2009) and Testing $\mathbf{A_t}$(2010-2014). Modified adjacency matrix is obtained using the method discussed in Subsection II-B. So, we have Training $\mathbf{M_s}$, Validation $\mathbf{M_v}$ and Testing $\mathbf{M_t}$. The validation will be used as the reference graph for learning methods from training. We use the spectral transformation of the graph kernels discussed in Section II. In real world the size of terrorist network is very large (as it is heterogeneous), giving large number of nodes. Working with large size of matrix is not preferred as we know that only some of the top eigenvalues can represent most of the information [26]. Therefore we use low rank eigenvalue decomposition [27] with a suitable rank $r << n$, where n is the total number of nodes in the graph. This approach limits the experimental iteration to the chosen rank, which is much lower than a full rank matrix. We set the rank to 30 which was found empirically.

Method: At first, low rank eigenvalue matrix decomposition with rank 30 is calculated for training and validation modified adjacency graph. Now, we have two data frames with eigenvalues, λ_s for training, λ_v for validation. As it is shown in [28] validation, if taken near to testing, gives better prediction for testing, we chose λ_v as the reference variable. Now using curve fitting we learn kernel's parameters and predict the λ_t. Matrix $\mathbf{M}^{'} = \mathbf{U_v} \times \mathbf{\Lambda_t} \times \mathbf{U_v^T}$, gives the global similarity score for all the node pairs, where $\mathbf{U_v}$ is the $n \times 30$ orthogonal eigenvector matrix for $\mathbf{M_v}$, $\mathbf{\Lambda}$ is 30×30 diagonal matrix contains λ_t on diagonal entries and $\mathbf{U_v^T}$ is the transpose of $\mathbf{U_v}$.

D. Data Preparation for Topic Modeling

Topic modeling differs SNA methods from other machine learning approach that it tries to utilize the hidden theme present in the network apart from the link information connecting various nodes. Through hidden theme captured by topic modeling, we are exploring relationship of terrorist group with country, city, target type, weapon used.

Method: We have have filtered out 10 columns from GTD datasets for our experiment as discussed in Subsection III . Each row of GTD corresponds to an event, and is treated as a document. Further, terms present in the column corresponding to each row like country code, city name etc) are treated as words of documents. We have used open source tool Gensim [29] for LDA implementation. Given the number of topic, LDA returns the topic-document relationship and topic-word relationship. Topic-word relationship can be viewed as bipartite graph as shown in figure 2. This bipartite graph contains topic and words as nodes. An edge in the above bipartite graph represents an association of topic and a word with scores obtained from LDA topic-word relationship. Now, for getting relationship between two entity like group and country, we use the following equation:

$$score(C_i,G_j)=\sum_{k=0}^{k=K}\left\{\frac{score(C_i,t_k)}{log(r(C_i,t_k))}+\frac{score(G_j,t_k)}{log(r(G_j,t_k))}\right\} \qquad (1)$$

where C_i refers to a country i, G_j refers to a group j, $score(C_i,t_k)$ represents the scores of country i in topic j. K is the total no of topic. $r(C_i,t_k)$ represents the rank of country i in topic k based on its contribution in the topic; similarly $r(G_j,t_k)$ represents the rank of group j in topic k. In the similar manner we find the association score between different city-group, target-group, and the weapon-group respectively.

We have experimented with different number of topic varying from 5 to 50 with a gap of 5. For each number topic , we have also experimented with considering different number of words for each topic ranging from 10 to 600 with a gap of 10 with selection threshold 0.0001. Empirically we found that topic 45 gives the best result for almost all cases. We are reporting only the peak results in Table V and Table VI because of size of the tables .

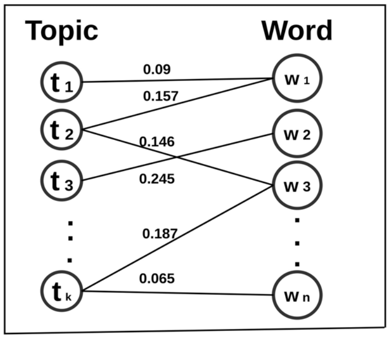

Figure 2. Bipartite graph between word and Topic obtained from LDA

Table IV
AVERAGE AUC SCORE FOR LINK PREDICTION ON ALL EDGES FOR WEIGHTED ADJACENCY MATRIX

Methods	Expo	Neu	PC4
Cn-gp	0.71	0.71	**0.73**
Cty-gp	0.85	0.85	**0.87**
Tar-gp	0.67	0.65	**0.73**
Weap-gp	0.79	0.78	**0.81**

E. Evaluation

We evaluate performance of all the link prediction models by finding their Area under ROC curve score. For this purpose, we generate 500000 edges randomly which are not existing in test graph. Evaluation was done in two steps for the test data. For the first case, we keep all the edges between year 2010 to 2014, while in second case, we remove those edges which appeared in training or validation. Motivation for dividing test data into these two forms is, we want to evaluate the models in terms of:

- Performance on seen as well as unseen connectivity (All Edges).
- Performance on unseen connectivity (Missing Edges).

The terrorist network, considered in this work generates a heterogeneous graph. It enables us to study relationship between different heterogeneous attributes. In this particular study, we have considered four pair of attributes relations; Country of attack vs Terrorist Group (**Cn-gp**) , City of attack vs terrorist group (**Cty-gp**), Target type vs terrorist group (**Tar-gp**), and Weapon type vs terrorist group (**Weap-gp**).

IV. RESULTS AND DISCUSSION

In this section, we present the performance for all the link prediction models described above, based on proximity based similarity measure, kernel and topic modeling . Following abbreviation was followed in all the tables presenting AUC score for the methods adopted. Cn: Country, Gp: Group, Cty: City, Tar: Target, Weap: Weapon, Expo: Exponential Kernel, Neu: Neumann Kernel, PC4: Path Counting Kernel with path length 4, CN: Common Neighbour, JC: Jaccard Coefficient, AA: Adamic-Adar Index, RA: Resource Allocation Index, TM: Topic Modeling.

Table V
AVERAGE AUC SCORE FOR LINK PREDICTION ON ALL EDGES

Methods	Expo	Neu	PC4	CN	JC	AA	RA	TM
Cn-gp	0.73	0.72	**0.78**	**0.78**	0.56	**0.78**	0.56	**0.82**
Cty-gp	0.80	0.79	**0.88**	**0.89**	0.42	**0.90**	0.42	**0.85**
Tar-gp	0.65	0.63	**0.71**	**0.74**	0.47	**0.75**	0.47	**0.68**
Weap-gp	0.76	0.74	**0.80**	**0.82**	0.33	**0.82**	0.33	**0.70**

Table VI
AVERAGE AUC SCORE FOR LINK PREDICTION ON MISSING EDGES

Methods	Expo	Neu	PC4	CN	JC	AA	RA	TM
Cn-gp	0.59	0.59	**0.68**	**0.67**	0.54	**0.68**	0.54	**0.71**
Cty-gp	0.75	0.75	**0.82**	**0.82**	0.45	**0.82**	0.45	**0.77**
Tar-gp	0.43	0.43	**0.50**	**0.53**	0.37	**0.53**	0.37	**0.57**
Weap-gp	0.46	0.46	**0.54**	**0.62**	0.30	**0.62**	0.30	**0.64**

Table V and Table VI, present the Average AUC score with the four evaluation criteria discussed in Subsection III-E for all edges and missing edges. It is evident that, Adamic-Adar method performs best among the models based on proximity based similarity measures, while Path Counting kernel performs best among graph kernels and topic modeling results are also competitive. From Table V, it is observed that, Topic Modeling performs best for the criteria, Cn-gp, while Adamic-Adar remains at the top for other three criteria. It is also noticed that Common Neighbour remains competitive with Adamic-Adar, throughout. We see in Table VI, that Topic Modeling performs best in the all criteria except Cty-gp. We observe that, unlike all edges, CN performs equivalently with AA for missing edges.

One of the reasons for best performance by AA and followed by CN, in almost all criteria for all edges, is the nature of the graph. This graph consists of set of cliques, which clearly shows that there will be more number of common neighbours between node pairs. JC and RA do not perform well and the reason is, scores are getting penalized by sum of the degree of common neighbours. Path Counting kernel gives score with a user defined path length k (4 for this work). So, it gives score by adding all the paths exist between node pairs, where k ∈ (2 to k). Therefore, it is observed that PC4, is quite competitive with respect to AA and CN. Exponential and Neumann kernels are based on equations, which are defined on power series of all length l, where l ∈ (2 to Total number of nodes) and penalized by the corresponding path lengths, which may be a reason for weaker performance than PC4.

Local proximity and kernel models exploit the structural information between nodes in a network based on the link structure of a network, while Topic modeling utilizes the hidden theme present in the network apart from the direct link information. Hence, Topic Modeling performs better than local proximity and kernels for missing edges, where direct links are absent.

All the above experiments are performed on the graph with equal weightage for all attributes. Further, we have examined the effect of different weights, chosen randomly for different features. It aims to verify the importance of a particular feature for performance on link prediction and reported in Table IV. For this task, experiments have been carried out using graph kernels as it is easy to modify the input adjacency matrix(for details refer Subsection II-B) in comparison to other link prediction methods discussed above. It is observed that, PC4 performs better than other graph kernels and results vary with changing weights.

V. Conclusion and Future Prospects

In this paper, we explored social network analysis methods such as, local similarity measures, graph kernels and topic modeling for predicting links in future on terrorist network namely, Global Terrorist Data. It is perceptible from the experimental results, that SNA methods are quite efficient for counter-terrorism with heterogeneous nature of a terrorist network. Influence based learning methods are successful on finding latent relation in heterogeneous networks. With this motivation, we investigate empirically the effect of influence for different attribute. It is found that, varying the importance or influence (in terms of weight), yields different prediction performance. This fact may be addressed as the future scope in network based counter-terrorism where idea would be to find the causal attributes importance first and then predicting the link in future.

References

[1] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, pp. 98–101, 2008.

[2] M. K. Sparrow, "The Application of Network Analysis to Criminal Intelligence: An Assessment of the Prospects," *Social Networks*, vol. 13, pp. 251–274, 1991.

[3] V. Krebs, "Mapping networks of terrorist cells," *CONNECTIONS*, vol. 24, no. 3, pp. 43–52, 2002.

[4] R. M. Medina, "Sna terrorism," *Secur J*, vol. 27, Feb 2014.

[5] J.-S. L. Kathleen Carley and D. Krackhardt, "Destabilizing terrorist networks," *CONNECTIONS*, vol. 24, no. 3, 2002.

[6] F. Spezzano, V. S. Subrahmanian, and A. Mannes, "Reshaping terrorist networks," *Commun. ACM*, vol. 57, no. 8, 2014.

[7] P. Jaccard, *Bulletin del la Société Vaudoise des Sciences Naturelles*, pp. 547–579.

[8] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, pp. 211–230, 2003.

[9] T. Zhou, L. Lü, and Y. Zhang, *The European Physical Journal B-Condensed Matter and Complex Systems*.

[10] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Phys. Rev. E*, Oct 2009.

[11] J. Kandola, J. Shawe-taylor, and N. Cristianini, "Learning semantic similarity," in *In NIPS*. MIT Press, 2003.

[12] R. Kondor and J. D. Lafferty, "Diffusion kernels on graphs and other discrete input spaces," in *(ICML 2002)*.

[13] D. E. Denning, *Networks and netwars: The future of terror, crime, and militancy*. National Defense Research Institute - RAND, 2001.

[14] J. A. Rodrguez, J. A. Rodrguez, and J. A. Rodrguez, "The march 11 th terrorist network: In its weakness lies its strength."

[15] M. Sageman, *Understanding Terror Networks*. University of Pennsylvania Press, 2004.

[16] K. M.Carley, J. Reminga, and N. Kamneva, "Destabilizing terrorist networks," in *NAACSOS*, 2003.

[17] J. Kunegis and A. Lommatzsch, "Learning spectral graph transformations for link prediction," ser. ICML '09. ACM.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, 2003.

[19] X. Wei and W. B. Croft, "Lda-based document models for ad-hoc retrieval," in *ACM SIGIR*. ACM, 2006.

[20] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[21] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *ACM SIGKDD*, 2011.

[22] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email," 2005.

[23] N. Pathak, C. DeLong, A. Banerjee, and K. Erickson, "Social topic models for community extraction," in *The 2nd SNA-KDD workshop*, vol. 8. Citeseer, 2008.

[24] G. L'huillier, S. A. Ríos, H. Alvarez, and F. Aguilera, "Topic-based social network analysis for virtual communities of interests in the dark web," in *ACM SIGKDD*, 2010.

[25] "National consortium for the study of terrorism and responses to terrorism (start). global terrorism database [data file]."

[26] Y. Qu, G. Ostrouchov, N. Samatova, and A. Geist, "Principal component analysis for dimension reduction in massive distributed data sets," in *ICDM*, 2002.

[27] D. Achlioptas and F. Mcsherry, "Fast computation of low-rank matrix approximations," *J. ACM*, vol. 54, no. 2.

[28] A. Anil, N. Sett, and S. R. Singh, "Modeling evolution of a social network using temporalgraph kernels," in *ACM SIGIR*.

[29] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.