INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

Link Prediction in Heterogeneous Information Networks: From Network Topology to Network Embedding

by

Akash Anil

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Department of Computer Science and Engineering

Under the supervision of

Dr. Sanasam Ranbir Singh

February 2020

Declaration of Authorship

I, Akash Anil, hereby confirm that:

- The work contained in this thesis is original and has been done by myself under the general supervision of my supervisor.
- This work has not been submitted to any other Institute for any degree or diploma.
- Whenever I have used materials (data, theoretical analysis, results) from other sources, I have given due credit to the authors/researchers by citing them in the text of the thesis and giving their details in the reference.
- Whenever I have quoted from the work of others, the source is always given.

Akash Anil
Research Scholar,
Department of CSE,
Indian Institute of Technology Guwahati,
Guwahati, Assam, INDIA 781039,
a.anil.iitg@gmail.com, a.anil@iitg.ac.in

Place: IIT Guwahati

Certificate

This is to certify that the thesis entitled **"Link Prediction in Heterogeneous Information Networks: From Network Topology to Network Embedding"** being submitted by **Mr. Akash Anil** to the department of *Computer science and Engineering, Indian Institute of Technology Guwahati*, is a record of bonafide research work under my supervision and is worthy of consideration for the award of the degree of Doctor of Philosophy of the Institute.

Dr. Sanasam Ranbir Singh
Department of CSE,
Indian Institute of Technology Guwahati,
Guwahati, Assam, INDIA 781039,
ranbir@iitg.ac.in

Place: IIT Guwahati

Acknowledgements

It gives me immense pleasure to thank each individual who supported directly or indirectly towards completion of my Ph.D journey. At first, I would like to thank my supervisor Dr. Sanasam Ranbir Singh for his exceptional and motivating guidance towards solving any real-life problem, may be related to Ph.D. or not. Moreover, his dedication towards his duties taught me many important lessons of life. I would always be indebted to him for several thought provoking ideas and constructive research discussions.

I would like to thank my Doctoral Committee members namely, Prof. SRM Prasanna, Dr. L. Boeing Singh, Dr. V. Vijay Saradhi, and Dr. Ashish Anand for their constructive suggestions towards shaping my research goals as well as the entire thesis. I have always gathered some fruitful insights whenever we had a discussion over any topics. Moreover, I was fortunate enough to learn various academic work-ethics by working as teaching assistant with Dr. V. Vijay Saradhi and Dr. Ashish Anand. I would like to thank Prof. Sukumar Nandi for his time and considerations on accepting the request to be a member of viva-voce evaluation committee. I was fortunate to work with Prof. Nandi in his team on a project. Although the working duration was little, I was able to learn many critical aspects and improve my technical management with the critical comments and productive suggestions from Prof. Nandi. Furthermore, my sincere thanks to Prof. S. V. Rao, the Head of the Department of Computer Science and Engineering and other faculty members for their direct and indirect support. I would like to express my gratitude to the thesis examiners namely, Prof. Ponnurangam Kumaraguru from IIIT Delhi and Prof. Tsuyoshi Murata from Tokyo Institute of Technology for a rigorous review and constructive suggestions.

I humbly thank to Mr. Raktajit Pathak, Mr. Nanu Alan Kachari, Mr. Bhriguraj Borah and all other institute's staffs for all the helps I borrowed towards making my journey smooth and productive. I specially thank Mr. Nanu Alan Kachari and Mr. Bhriguraj Borah for their extreme dedications towards managing efficient computing facilities at the department. My thesis would not have been completed without their timely support. Furthermore, I would like to thank IIT Guwahati administration for providing on-campus hostel facility. From the core of my heart, I would like to thank the mess staffs, canteen staffs, security personals, and housekeeping staffs for making my stay memorable and smooth.

Having good friends is always a blessing. Fortunately, I have a very large set of good and close friends with whom I have spent very quality time. I am privileged to mention Manohar Singh Gour, Subhrendu Chattopadhyay, Kundan Kumar, and Durgesh Kumar as four long time peers who supported my entire journey in several perspectives. I had the privilege of having a very helpful and supportive seniors (as friends), namely Dr. Niladri Sett, Dr. Sounak Chakraborty, Dr. Satish Kumar, Dr. Sibaji Gaj, Kunwer Mrityunjay Singh, Awnish Kumar, Rahul Gangopadhyay, Dr. Rajesh D, and Mausam Handique. I would like to mention Dr. Niladri Sett and Dr. Satish Kumar especially for their critical comments and road-maps which made my Ph.D. journey more smoother and efficient. I really feel privileged to have many joyous moments with Brijesh, Prasen, Abhishek, Ghalib, Surajit, Saptarsi, Sunil, Swarup, Hema, Dipojjwal, Akshay, Pawan, Shashi, Agyapal, Swati, Kamal, Adi, Jiten Da and Bala. My stay at IIT Guwahati was made more pleasant by having many good memories with friends from OSINT lab like Neelakshi, Hemanta, Gyanendro, Sujit, Lenin, Bornali, Mala, Pankaj, Tonmoya, Deepen, Anupam, Jubanjan, Anasua, Rajib Sir, Piyush, Pardeep, Neelesh, Nitesh, Ranjan, Rakesh, Rajlakshmi, Akhilesh, Debashish and many more.

During my Ph.D. Journey, I was fortunate to work with many creative minds of IIT Guwahati. Some of them are Durgesh, Ranjan, Nitesh, Rakesh, Uppindar, Shubham, Piyush, and Gyanendro. Moreover, I have also worked with Sandeep and Ajay from NIT Silchar when they were interns at IIT Guwahati. Our countless discussions and dedications paved a fruitful way to shape my Ph.D. Furthermore, I would like to thank all the anonymous reviewers of my papers & thesis as well as friends I forgot to mention here.

The success of my Ph.D. is due to the constant support and motivations from my family members. From the core of my heart, I would like to thank my parents for believing in my quest and supporting me blindly in all the ups and downs. I would like to thank my elder brother and sister in law for all the encouragements towards my goal. Having kids around is always joyful. I would like to thank my nieces, namely, Drishti and Divya for all their cute discussions over phone. Furthermore, I would like to thank the new member of our family, namely, Meenakshi for all the support, care, and togetherness. Finally, I would like to express my gratitude to all the friends and relatives for helping my family in various needs when I was away from home.

Abstract

Modeling real-world systems using complex network analysis has become a popular approach in the last two decades. A complex network is loosely divided into four types of networks, namely (i) Social Network, (ii) Information Network, (iii) Technological Network, and (iv) Biological Network. However, most real-world networks can be represented as Information Network. Majority of the previous literature over information networks consider homogeneous network representation (singular types of nodes and relations) e.g., Citation network, World Wide Web, etc. However, it has recently been realized that Heterogeneous Information Network (HIN) that consists of multiple types of nodes and relations is a better representation for real-world physical systems. For example, a HIN representing a Citation network by considering node types, such as Author, Paper, Venue, etc. and their corresponding relations, captures rich semantics in comparison to the homogeneous Citation network (considering only the papers as node and citation as relation). Motivated with this, our objective is to leverage Heterogeneous Information Network representation in modeling evolution of a given system by solving link prediction problem. In particular, the major contributions of this thesis towards link prediction can be divided into three types of approaches; (i) topology-based, (ii) graph kernel-based and (iii) network embedding-based. For topology-based methods, we adapt the state-of-the-art common neighbor-based local similarity measures to heterogeneous information network. For graph kernels-based methods we propose a generalized heterogeneous framework for state-of-the-art spectral graph kernels. Furthermore, in network embedding-based methods, we exploit k-hop random-walk to generate node neighborhood for training the model. From previous studies, it is evident that majority of the complex networks are susceptible to the exogenous information (e.g., news and social media) apart from endogenous information (e.g., network characteristics such as clustering coefficients, degree distribution, etc.). Therefore, we study the effects of exogenous information such as news media, temporal dynamics of the underlying network on the proposed topology-based heterogeneous similarity measures. We observe that incorporating exogenous information helps in boosting performance of the link prediction. Further, majority of the studies on link prediction using network embedding are based on meta paths, we critically analyze the efficacy of meta path-based methods over link prediction and node classification tasks. We observe that heterogeneous network embedding cannot be generalized and meta path-based embeddings are task-specific. As most of the heterogeneous information network having different

number of instances for different types of nodes and relations, class imbalance is inherent in such networks. Therefore, this thesis further studies the effects of class imbalance in heterogeneous network embedding. It is observed that selecting appropriate node types along with addressing class imbalance in heterogeneous information network is an important pre-requisite for efficient network embedding.

Contents

List of Figures

List of Tables

List of Symbols

G	Graph/Network
V	Set of nodes
E	Set of edges
$\mathcal{V}$	Set of node types
$\mathcal{E}$	Set of edge types
$\|X\|$	Cardinality of the set X
$\mathcal{P}$	Meta-path
N	Total Number of nodes in the network
$n(x)$	Set of neighbor nodes for x
$S_{CN}(x, y)$	Likelihood score for nodes x and y by common neighbor
$S_{JC}(x, y)$	Likelihood score for nodes x and y by Jaccard Coefficient
$S_{AA}(x, y)$	Likelihood score for nodes x and y by Adamic Adar
$S_{RA}(x, y)$	Likelihood score for nodes x and y by Resource Allocation
$\mathbb{R}$	Set of Real Numbers
$\mathcal{N}_x$	Set of node sequences for node x
$P(x)$	Probability of x
RW-k	Random Walk capturing k-hop proximity
π	Transition Probability
$\mathcal{C}_{min}$	Set of minority class nodes
$\mathcal{C}_{maj}$	Set of majority class nodes

Abbreviations

HIN	Heterogeneous Information Network
HTAN	Heterogeneous Terrorist Attack Network
WWW	World Wide Web
GTD	Global Terrorism Database
CN	Common Neighbor
JC	Jaccard Coefficient
AA	Adamic-Adar Index
RA	Resource Allocation
PC	Path Counting
EXP	Exponential
NEU	Neumann

Chapter 1

Introduction

Majority of the real-world systems such as computer communication, biological network, terrorist attack, bibliography, transportation etc. are defined by multiple interacting objects [2]. Further, any such physical system can be represented as a complex network consisting of objects as nodes and interaction types as relations or links [3]. In network science, such networks are often referred to as `Information Network` as they represent a particular information or data [4]. Some of the examples of information networks are (i) Citation network in bibliographic data having paper as nodes and citation relation as links, (ii) Co-authorship network in bibliographic data having authors as nodes and co-authorship relation as links, (iii) World Wide Web where web-pages are the nodes and links are defined by hyperlinks, (iv) Social networks where users are the nodes and friendship as the relations, etc. For the systems where feature vector generation is not trivial, an information network offers elegant ways of solving various machine learning problems such as classification [5], clustering [6, 7], link prediction [8], ranking [9, 10], recommendation [11, 12], etc. by exploiting graph theory and network-science-based methods.

In past two decades, graph-theoretic and network-based approaches have garnered a considerable maturity. However, majority of the previous studies consider information networks with singular type of nodes and relations, i.e. homogeneous information network [13]. As real-world physical systems often have multiple

types of objects and interactions, a homogeneous information network may be a lossy representation. For example, a bibliographic data consisting of multiple objects such as Author, Paper, Venue, Topic, etc. related to each other by different semantics may not be represented well using a homogeneous information network. To create a lossless representation for the given system, recent studies [2, 4, 14, 15, 16, 17, 18, 19] consider Heterogeneous Information Network (HIN) where multiple types of objects and interactions are represented by multiple types of nodes and relations. A HIN compared to homogeneous information network, captures rich and complex structural characteristics with multiple semantics [13]. Thus, there is a surge in devising network-mining models having capability of exploiting the heterogeneous characteristics of real-world data while solving various network mining tasks.

Incorporating multiple types of nodes and relations in HIN results in multiple types of paths between same pair of nodes. In literature, these paths are referred to as meta-paths [20]. A majority of the previous studies over HIN exploit different meta-paths for solving various network-mining problems [21, 22, 23, 24, 25]. In general, these studies generate network's features using predefined meta-paths-based on some heuristics. Thereafter, learning models are trained using some of the classification or regression frameworks. It is evident from previous studies that meta path-based features help in various prediction tasks over HIN. However, selecting best meta-path becomes a major bottleneck when the underlying HIN has large number of node and relation types. Thus, finding an optimal meta-path is a critical issue in HIN analysis based on meta-path-based features.

Although a HIN represents real-world system in a natural way and offers rich source of information, exploiting HIN for solving various real-world problems is a challenging task due to following reasons [13]:

1. Since there is no standard way of constructing HIN, there might be multiple HINs representing the same system.

2. As HINs consider multi-typed nodes and edges, an unique proximity estimate between two nodes cannot be generalized. It is because there may exist many paths between these nodes representing different semantics.

3. Previously proposed methods based on meta-paths for exploring HIN, may be task-specific or sensitive to different meta-paths. Moreover, finding best meta-path is not trivial for majority of the HINs having large number of node and edge classes.

4. As HINs permit different number of instances for different types of nodes, class imbalance can be a critical issue while solving various problems like ranking, link prediction, etc.

Motivated from the above challenges, this thesis re-visits the heterogeneous information network representation for real-world datasets and attempts to study the applicability of state-of-the-art methods as well as propose new methodologies for link prediction in HIN.

1.1 Commonly Studied Problems over HIN

A HIN is a special case of complex network built over real-world system having multiple types of nodes and relations [3, 13]. Thus, a most of the HINs exhibit similar characteristics as shown by various complex networks such as (i) a heavy tail in the degree distribution, (ii) a high clustering coefficient, (iii) community structure, and (iv) hierarchical structure [26]. Consequently, a majority of the commonly studied problems in complex network can also be studied over HINs. The popular problems studied over complex networks are as follow:

1. **Centrality Estimation [27]** is the process of finding central nodes in the underlying network based on their importance, influence, and prestige. Centrality estimation helps in finding important objects in a system which control the spread of information and relative influence among various objects.

Some of the state-of-the-art centrality measures are degree [28], betweenness [29], closeness [30], eigenvector [31], PageRank [9], etc.

2. **Link Prediction [32, 33, 8, 34]** in a network refers to predicting relations in future or missing relations by using the existing network structure. Link prediction helps in various applications such as recommending new friends on online social networks, ranking web-pages highly related to the search query on search engines, etc. Some of the state-of-the-art link prediction methods are Common Neighbor [8], Jaccard Coefficient [35], Adamic Adar Index [36] Resource Allocation [37], Preferential Attachment [8], Katz Index [8], etc.

3. **Information Diffusion** studies the factors playing important role in the evolution of opinions or status of a node in the underlying network. Moreover, information diffusion has wide range of applications in real-world networks such as forecasting election result [38], localization of natural disasters [39], targeted advertising and personalized marketing [40, 41, 42], etc.

4. **Community Detection** deals with identifying groups or communities in the underlying network such that nodes inside the community are densely connected whereas nodes outside the community have sparse connections [43, 44]. Community structure is an inherent characteristics of the real-world networks and helps in various applications such as rumor spreading [45, 46], epidemic spreading [47], etc.

5. **Network Re-construction** focuses on devising statistical and mathematical models to re-construct a complex network from the observed sample of network [48]. The main challenges of network re-construction are the observed sample usually exhibit limited characteristics and may have noise. Some of the popular approaches of network re-construction are collective dynamics [48, 49, 50, 51, 52], stochastic analysis [53, 54], and compressive sensing [55, 56, 57, 58].

The above categorization of the problems are based on specific intuition and motivation. However, all of them harness the proximity among nodes in the underlying

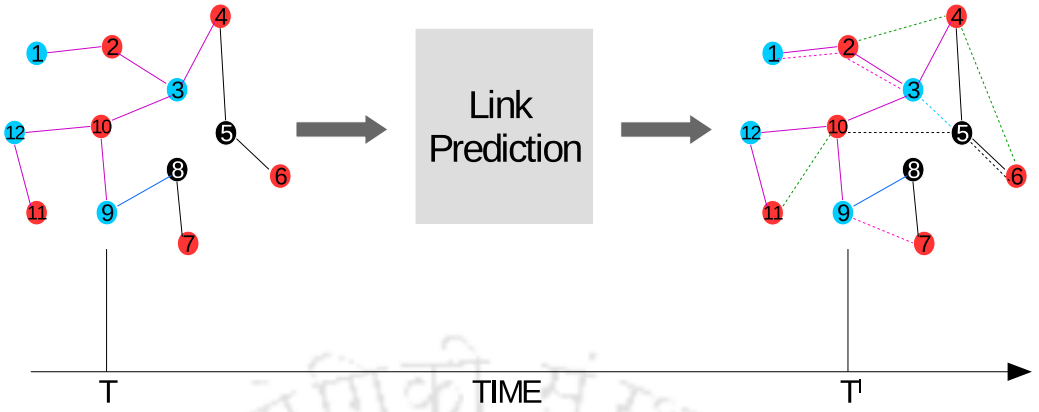


FIGURE 1.1: Link prediction in future over heterogeneous information network, dotted lines represent repeated or new relations in future.

network to solve the particular objective. In this thesis, we focus on link prediction as an application and devise novel frameworks which can incorporate the HIN characteristics while capturing proximity between two nodes.

1.2 Problem Statement

As discussed above, link prediction in a given network can be defined in two perspectives:

1. Use the observed network structure and predict the likelihood of missing relations [32, 33], e.g. given a synonymy network where words are represented as nodes and edges are defined by synonymy relation, predict the likelihood of having synonymy relations among other disconnected nodes.

2. Use the observed network structure and predict the likelihood of relations in future [8, 34], e.g. given a co-authorship network where authors are represented as nodes and edge between two nodes are defined by co-authorship relation, predict future (repeated or new) co-authorship relation.

While both of the above perspectives for predicting links have important applications in real-world problems, the second perspective is more suitable for the networks which are *longitudinal* (evolve with time) in nature, e.g. social networks [59]. As this thesis explores longitudinal heterogeneous information networks such as terrorist attack network, bibliographic network, etc., we consider the second perspective of link prediction. Thus, this thesis defines the problem statement as follows: *Given a heterogeneous information network at time T, predict relations in future time $T^{'}$ where $T < T^{'}$.*

Figure 1.1 presents the graphical illustration for link prediction task. The input to the link prediction model is a heterogeneous information network (at time T) with multiple types of nodes and edges (distinguished by different colors). Thereafter, the link prediction framework outputs a heterogeneous information network with edges in future i.e. at $T^{'}$ (represented as dotted lines). It must be noted that the future relations may be completely new links or repeated links. For example, in Figure 1.1, relations `1-2, 2-3, 5-6` are repeated relations in future whereas other dotted lines represent new relations appeared in future.

As discussed above, majority of the previous studies exploit meta-path-based features for link prediction in HIN. However, finding best meta-path is itself a research problem. Thus, in a different direction this thesis attempts to explore HIN for link prediction problem that do not require any explicit meta-path. Further, the major aspects of HIN studied in this thesis are as follow:

1. **Node Importance:** As a HIN considers multiple types of nodes, different types of nodes may have different influence on formation of future relations. Thus, this thesis formulates heterogeneous versions of state-of-the-art link prediction methods capable of incorporating node importance while estimating likelihood of link existence.

2. **Exogenous Information:** Most real-world networks are susceptible to exogenous information apart from endogenous network structure. For example, an user on twitter may post a message by sharing an URL (exogenous) or

re-tweeting a tweet posted by an user from her/his following list (endogenous). Thus, this thesis incorporates exogenous information such as news media and temporal dynamics of the underlying network while predicting future links.

3. **Network's Latent Representation:** Although, majority of the state-of-the-art link prediction methods exploit the explicit link structure of the underlying network, two nodes may show similar characteristics (e.g. hubs) even though they are not connected via any path. Thus, latent representation of the network or network embedding may be useful to generate automatic node features. Thus, this thesis further studies and propose novel network embedding method for HIN embedding.

4. **Class Imbalance:** As HINs consider multiple types of nodes and relations and there is no control over number of instances of particular node or relation types, class imbalance is intrinsic in HINs. Thus, this thesis studies the effects of class imbalance in HIN embedding subjected to different network-mining problems.

1.3 Contribution

This section briefly summarizes the contributions made in this thesis.

1.3.1 Link prediction in HIN

This work focuses on formulating heterogeneous versions of local similarity measures and spectral graph kernels. In particular, we adapt the state-of-the-art local similarity measures capturing triadic closure (proximity of path length two) and path-based spectral graph kernels to predict links in HIN. We formulate the heterogeneous versions of the local similarity measures and spectral graph kernels by incorporating node importance. The proposed heterogeneous frameworks are easy

to generalize for incorporating different types of node importance while predicting future links.

1.3.2 Link Prediction Using Exogenous Factors as Node Importance

As discussed above, majority of the applications over real-world networks are affected by exogenous information such as news media, online blogs, etc. Therefore, the objective of this work is to incorporate effects of exogenous information (e.g. news media and network's temporal dynamics) for predicting future relations in a Heterogeneous Terrorist Attack Network (HTAN). We exploit the above proposed heterogeneous local similarity measures for link prediction task. As these heterogeneous local similarity measures require node importance, we estimate centrality of the nodes using Personalized PageRank (PPR). We exploit multiple exogenous information as personalization parameters to PPR and generate node weighted HTAN. This node weighted HTAN is further used by heterogeneous similarity measures to estimate the likelihood of future relations among different nodes.

1.3.3 Network Embedding for HIN

The analytical approaches exploiting only the network topology may not capture several characteristics of the underlying network such as structural role, node equivalence, etc. Thus, many of the studies in past explored network representation in hidden space or network embedding framework that generates feature vectors for the nodes such that two nodes having similar characteristics are close in features space.

This work studies the problem of network embedding in HIN. In particular, we focus on a critical analysis of recently proposed meta-path-based frameworks for

HIN embedding. Moreover, we study the applicability of different types of meta-paths for HIN embedding over problems of diverse nature, i.e. link prediction, classification, etc.

1.3.4 HIN Embedding using k-hop Random Walks

From the above study, we observe that generalization of meta-path-based HIN embedding is a critical issue. Moreover, selecting best meta-path is not a trivial problem. Therefore, this study proposes a novel HIN embedding framework that do not require any meta-path explicitly. In particular, the proposed model exploits k-hop random walks to capture network neighborhood and using skip-gram-based neural network model we generate the node features. We exploit the node embeddings for Co-authorship prediction task in heterogeneous bibliographic network. Furthermore, the performance of the proposed model is compared with recent state-of-the-art network embedding methodologies. We observe that our proposed network embedding model outperforms the baselines in a majority of the cases.

1.3.5 Effect of Class Imbalance in HIN Embedding

HIN incorporates multiple types of nodes and relations. Further, there may be irregular number of instances in different types of nodes and relations. Thus, class imbalance is an intrinsic characteristics for a majority of the HINs. Previous works on HIN embedding use meta-paths for addressing the class imbalance issue. However, none of these works study the effects of class imbalance in detail. Therefore, this work presents a comprehensive study focused on investigating the effects of class imbalance in HIN over HIN embedding. We study the performance of state-of-the-art network embedding models subjected to HINs having different level of class imbalance. We observe that for better HIN embedding, node types selection is one of the important pre-requisite along with addressing class imbalance.

1.4 Outline of the Thesis

The outline of this thesis is as follows.

- Chapter 2: Background
  - Discusses the important pre-requisites which are used in the thesis.
- Chapter 3: Link Prediction in HIN
  - Proposes heterogeneous frameworks of local similarity measures and spectral graph kernels for link prediction in HIN.
- Chapter 4: Link Prediction Using Exogenous Factors as Node Importance
  - Exploits Personalize PageRank to incorporate exogenous information as node importance for link prediction task.
- Chapter 5: Network Embedding for HIN
  - Investigates effects of different meta-paths over HIN embedding.
- Chapter 6: HIN Embedding using k-hop Random Walks
  - Proposes a novel model for HIN embedding based on $k-$hop random walks.
- Chapter 7: Effect of Class Imbalance on HIN Embedding
  - Investigates the effects of class imbalance for solving various network-based problems using HIN embedding.
- Chapter 8: Conclusion and Future Work
  - Concludes the thesis and discusses the future directions.

Chapter 2

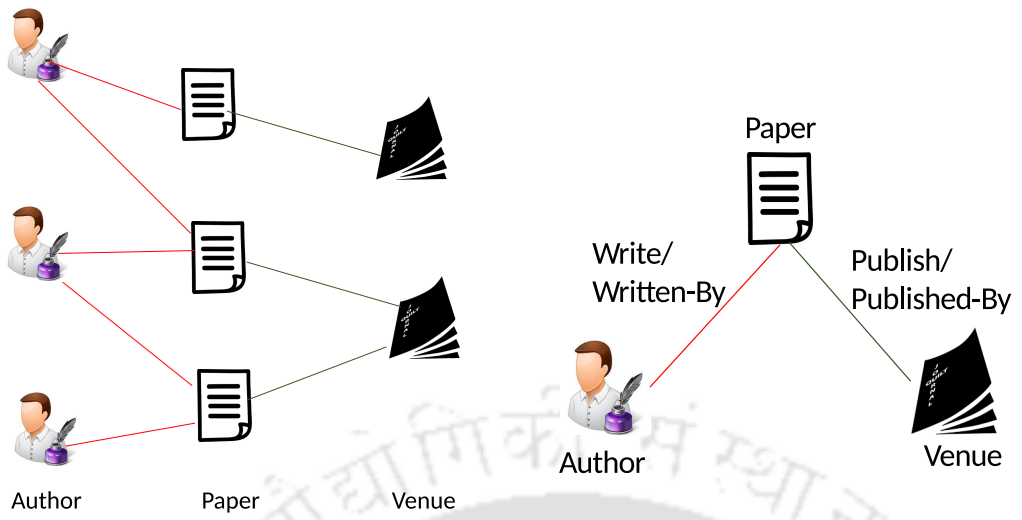
Background

This thesis is mainly built over two independent approaches for link prediction in heterogeneous information network. The first approach is to exploit the network topology (e.g. topological similarity measures) while second focuses on representing the network in latent representation (i.e. network embedding). Therefore, this chapter briefly discusses the important pre-requisites for a better understanding of the subsequent chapters.

The subsequent chapters present the major contributions made in this thesis. We discuss the related studies in regards to individual contributions in the corresponding chapters. Thus, this thesis does not consists of an independent chapter for literature survey.

2.1 Information Network

Information network can be referred to as a complex network consisting of real-world objects (e.g. people, articles, computers, etc.) connected to each other via complex relations (e.g. friendship, authorship, peers, etc.). Information networks are also refrred to as Knowledge Networks [3]. One of the mostly studied information network is citation network. In citation network, academic papers are treated as nodes and relations are defined by directed edges demonstrating paper



(a) HIN Instance (b) HIN Schema

FIGURE 2.1: An example of bibliographic heterogeneous information network.

citing previous papers [60]. A citation network is acyclic as all the papers cite papers published in the past. Another important example of information network is World Wide Web (WWW) where web-pages are the nodes and two nodes are connected if there is hyperlink between them [61]. Unlike citation network, WWW network can be cyclic as there is no restriction on building hyperlinks among different web-pages. Other examples of information networks are *peer to peer networks of computer [62, 63, 64], word network in thesaurus [65, 66, 67, 68], preference network over movie database [69, 70, 71], etc.* Although, the field of information network is quite mature today, yet it is an active area of research.

Definition 2.1 (Information Network [4, 72])**.** An Information Network can be defined as a six-tuple: $< V, E, \mathcal{V}, \mathcal{E}, \phi, \psi >$ where, V is a set of nodes, E is a set of edges, $\mathcal{V}$ is a set of node classes, and $\mathcal{E}$ is a set of edge classes. Any node $v \in V$ belongs to one of the classes in $\mathcal{V}$ and any edge $e \in E$ belongs to one of the classes in $\mathcal{E}$. Further, $\phi : V \rightarrow \mathcal{V}$ is a node class mapping, and $\psi : E \rightarrow \mathcal{E}$ is a link class mapping. Furthermore, an information network is homogeneous if $|\mathcal{V}| = 1$ and $|\mathcal{E}| = 1$. Otherwise, the given network is Heterogeneous Information Network (HIN).

Figure 2.1(a) shows an instance of a bibliographic HIN consisting `Author`, `Paper`,

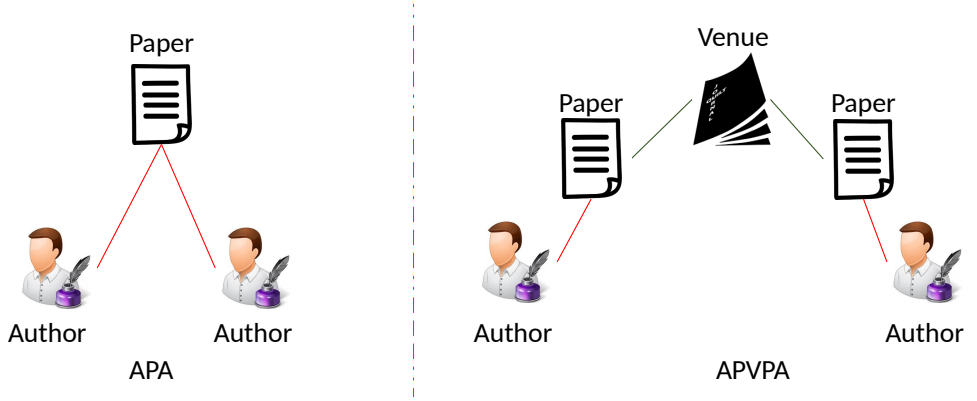


FIGURE 2.2: Two types of meta-paths in the bibliographic heterogeneous information network shown in Figure 2.1.

and `Venue` as nodes. Further, this HIN follows the network schema shown in Figure 2.1(b) and consists of two types of relations, namely (i) Authors `Write` Papers or Papers `Written-By` Authors and (ii) Papers `Published-By` Venues or Venues `Publish` Papers.

2.2 Meta-path

From the schematic illustration of HIN presented in Figure 2.1, it is evident that a HIN consists of multiple types of nodes and edges. Thus, there may exist multiple types of paths among nodes. Moreover, these paths shall represent proximity of different semantics between the given pair of nodes. In previous studies, these paths have been referred to as Meta-path which is defined below [73].

Definition 2.2 (Meta-path)**.** For a given HIN $G(V, E, \mathcal{V}, \mathcal{E}, \phi, \psi)$, a meta-path $\mathcal{P}_{(\mathcal{V}_1,\mathcal{V}_l)}$ can be defined as a sequence of edge types between node types $\mathcal{V}_1$ and $\mathcal{V}_l$ which is represented as $\mathcal{V}_1 \xrightarrow{\mathcal{E}_1} \mathcal{V}_2 \xrightarrow{\mathcal{E}_2} \cdots \xrightarrow{\mathcal{E}_{l-1}} \mathcal{V}_l$.

The length of the meta-path is given by the number of edge types it considers. Further, a symmetric meta-path is capable of representing a relation between

two nodes. Figure 2.2 shows two types of meta-paths, namely Author-Paper-Author (APA) and Author-Paper-Venue-Paper-Author (APVPA) over the HIN schema presented in Figure 2.1(b). These two meta-paths capture proximity of different semantics between the given pair of nodes. For example, APA captures co-authorship relation whereas APVPA captures proximity between two authors publishing papers at the same venue.

2.3 Local Similarity Measures

This thesis exploits state-of-the-art local similarity measures capturing local proximity of path length two (triadic closure) for predicting future links between two nodes. In particular, we consider four local similarity measures, namely *(i) Common Neighbor [8], (ii) Jaccard Coefficient [35], (iii) Adamic-Adar index [36], and (iv) Resource Allocation [37].*

1. **Common Neighbor (CN):** It is one of the simplest local similarity measure which estimates the similarity between two nodes by counting number of common neighbors they have. In other words it counts the number of paths having length two between the given pair of nodes. Two nodes having large number of common neighbors are regarded more similar. Let $n(v)$ gives a set consisting of all the neighbors of node v then common neighbor score between two nodes x and y can be defined as follows:

$$S_{CN}(x, y) = |n(x) \cap n(y)| \tag{2.1}$$

2. **Jaccard Coefficient (JC):** The main intuition behind Jaccard Coefficient is that *two nodes shall have less similarity if these nodes are similar to many other nodes.* In a network, the immediate neighbors of a particular nodes are often considered as the most similar nodes. Thus, Jaccard Coefficient is estimated using a normalization factor along with the CN similarity score. Formally, JC is defined as follows:

$$S_{JC}(x,y) = \frac{|n(x) \cap n(y)|}{|n(x) \cup n(y)|} \tag{2.2}$$

3. **Adamic-Adar (AA):** The above-defined CN and JC similarity measures treat every common neighbors equally. However, in social networks, it is often visible that a common neighbor having high connections (degree) do not contribute much to the similarity between its neighbors. In different direction, Adamic-Adar index treats each common neighbors differently. In particular, a common neighbor having higher connections contributes less whereas common neighbor with less connections contributes high to the Adamic-Adar similarity score. Adamic-Adar similarity measure is defined as follows:

$$S_{AA}(x,y) = \sum_{z \in n(x) \cap n(y)} \frac{1}{log(|n(z)|)} \tag{2.3}$$

4. **Resource Allocation (RA):** This index is based on the motivation from resource allocation dynamics in complex networks [74] which states that every resource will equally be shared among its neighbors. Moreover, RA formulation is similar to the AA, i.e. common neighbor nodes having high connection contribute less to similarity score and vice-versa. Thus, RA is defined as follows:

$$S_{RA}(x,y) = \sum_{z \in n(x) \cap n(y)} \frac{1}{|n(z)|} \tag{2.4}$$

2.4 Spectral Graph Kernels

Given a graph $G(V, E)$, a graph kernel is a measure of similarity between a pair of nodes. Graph kernels have been used for several network-science problems such as link prediction [75, 76, 77, 78, 79, 80], network growth [81, 82], community detection [83], etc. A graph kernel follows properties like symmetry and positive semi-definite. In previous studies, many graph kernels have been studied using spectral

graph theory [84, 85]. In particular, the spectral transformation of the given graph kernel decomposes an adjacency or Laplacian representation of the given graph. This thesis focuses on leveraging adjacency-based spectral graph kernels, namely (i) Path Counting [86], (ii) Exponential [87], and (iii) Neumann [87].

2.4.1 Graph Kernels and their spectral transformations

Spectral graph theory is a popular method to study graph properties [84, 85], which exploits decomposition of adjacency or Laplacian matrix of a graph. Commonly used decomposition method for square matrix is eigenvalue decomposition (EVD). Considering a square adjacency matrix $\mathbf{A}$ of a graph, EVD can be used to decompose $\mathbf{A}$ as follows.

$$\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^T \tag{2.5}$$

where Λ is a diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{A}$ in decreasing order and $\mathbf{U}$ is a matrix whose i^{th} column is an eigenvector corresponding to the i^{th} eigenvalue of $\mathbf{A}$ i.e., Λ_{ii}. Considering Equation (2.5), spectral transformation of a graph kernel $K(\mathbf{A})$ is defined as

$$K(\mathbf{A}) = K(\mathbf{U}\Lambda\mathbf{U}^T) = \mathbf{U}F(\Lambda)\mathbf{U}^T \tag{2.6}$$

for some kernel function $F(\Lambda)$. Similar to [88], this thesis also assumes that eigenvectors are time invariant. The function $F(\Lambda)$ is defined by a set of real valued function$f(\lambda)$, where λ is a vector containing eigenvalues of the adjacency matrix $\mathbf{A}$.

1. **Path Counting (PC):** Path counting is a graph growth model which is defined by sum of all possible paths upto length k between the participating nodes [86]. It can be expressed as $\sum_{i=0}^{k} \mathbf{A}^k$. Path counting model can be transformed into a graph kernel by using only the top k positive eigenvalues of the matrix $\mathbf{A}$. Further, $\mathbf{A}^k$ can be expressed as follows.

$$\mathbf{A}^k = \mathbf{U}\Lambda^k\mathbf{U}^T \tag{2.7}$$

From Equation (2.6), spectral path counting kernel can be defined as

$$PC(\sum_{i=0}^{k}\mathbf{A}^k) = \mathbf{U}F(\sum_{i=0}^{k}\Lambda^k)\mathbf{U}^T$$

and $F(\sum_{i=0}^{k}\Lambda^k)$ is given by function $f(\lambda) = \sum_{i=0}^{k}\alpha_i\lambda^k$ where, $\alpha \neq 0$.

2. **Exponential Kernel (EXP):** Exponential kernel [87] estimates exponential of the adjacency matrix $\mathbf{A}$ which can be defined as follows.

$$EXP(\mathbf{A}) = \exp(\mathbf{A}) = \sum_{i=0}^{\infty}\frac{\alpha^i}{i!}\mathbf{A}^i = e^{\alpha\mathbf{A}} \tag{2.8}$$

Further, From Equation (2.6), exponential kernel using spectral transformation can be defined as

$$EXP(e^{\alpha\mathbf{A}}) = \mathbf{U}F(e^{\alpha\Lambda})\mathbf{U}^T$$

and $F(e^{\alpha\Lambda})$ is given by function $f(\lambda) = e^{\alpha\lambda}$ where,

3. **Neumann Kernel (NEU):** The Neumann kernel [87] of a node pair is defined as follow.

$$NEU(\mathbf{A}) = (\mathbf{I} - \alpha\mathbf{A})^{-1} \tag{2.9}$$

where, α is chosen such as $\alpha^{-1} > \lambda_1$ (principal eigenvalue of $\mathbf{A}$). From Equation (2.6), Applying spectral transformation, we get spectral Neumann kernel as

$$NEU(\mathbf{A}) = [\mathbf{U}F(\mathbf{I} - \alpha\Lambda)\mathbf{U}^T)]^{-1}$$

and $F(\mathbf{I} - \alpha\Lambda)$ is given by $f(\lambda) = \frac{1}{1-\alpha\lambda}$, where, $\alpha \neq 0$.

2.5 Centrality Measures

This Section briefly discusses the different centrality measures used in this study. As this thesis exploits undirected HINs, we limit the scope of definitions to undirected networks only.

1. **Degree Centrality:** This is one of the simplest centrality measures that estimates the centrality of a node based on the number of connections it has with other nodes [28]. Suppose in a network $G(V, E)$, x_i gives the centrality score for a node $v_i \in V$ and $\mathbf{A}$ be the adjacency representation of the underlying network, then Degree centrality can be defined as follows:

$$x_i = \sum_j A_{ij} \tag{2.10}$$

where $A_{ij} = 1$ if there is an edge between nodes v_i and v_j, otherwise 0.

2. **Betweenness Centrality:** Betweenness centrality for a node is characterized by the number of times a node appears in the shortest paths between all possible pairs of nodes in the network. Suppose, $v_s \in V$ and $v_t \in V$ be any pair of nodes in a network, then Betweenness centrality of node $v_i \in V$ can be defined as follow [29]:

$$x_i = \sum_{st} \frac{N_{st}^i}{g_{st}}, \tag{2.11}$$

where, N_{st}^i is the total number of times v_i appears on shortest paths between v_s and v_t, g_{st} is the total number of shortest paths available between v_s and v_t.

3. **Closeness Centrality:** This centrality measures characterizes the central nodes by estimating mean distance of the given node to other nodes in the network. A node with lowest mean distance from other nodes is treated as most central node. Suppose $d_{i,j}$ is the length of shortest path between nodes $v_i \in V$ and $v_j \in V$ and N be the total number of nodes, then the mean

distance of node v_i from other nodes v_j in the network can be estimated as: $l_i = \frac{1}{N}\sum_j d_{ij}$. Now, the Closeness centrality can be defined as follow [30]:

$$x_i = \frac{1}{l_i} = \frac{N}{\sum_j d_{ij}} \tag{2.12}$$

4. **Eigenvector Centrality:** Eigenvector centrality is seen as a natural extension to degree centrality. However, unlike degree centrality which awards one centrality score to the node for having a neighbor, eigenvector centrality awards a score to the node which is proportional to the sum of scores of its neighbors. Eigenvector centrality can be defined as follow [31]:

$$x_i = \sum_{x_j \in n(x_i)} A_{ij} x_j, \tag{2.13}$$

5. **PageRank:** PageRank is a random walk based centrality measure. PageRank was first used by Google search engine to measure the centrality estimates of the web-pages in WWW information network. PageRank awards centrality score to a web-page by counting number and quality of hyperlinks it has. In other words, PageRank awards importance to a node based on the number and the importance of its neighbors. Let $\mathbf{M}$ be a row stochastic matrix representing the underlying network such that $\sum_{j=1}^{N} M_{ij} = 1$ for $i = 1, 2, 3, \cdots, N$, then PageRank can be defined as follow [9]:

$$x_i = d\sum_{j=1}^{N} x_j M_{ji} + (1-d)\frac{1}{N} \tag{2.14}$$

where $d \in [0, 1]$ is the damping parameter.

2.6 Network Embedding

Network Embedding refers to representing a network in terms of low dimensional vectors corresponding to node, edge, subgraph, etc. which can further be used for various network-mining tasks [89]. The scope of this thesis is to generate node

embedding. Node embedding has been carried out in several ways such as, *(i) representing normalized adjacency or Laplacian matrix to lower dimensions using non-linear dimensionality reduction techniques [90, 91, 92], (ii) eigenvalue decomposition [93], (iii) singular value decomposition [94], and (iv) neural network-based representation learning [95, 96, 1, 97, 98, 99, 100, 101, 102, 103].* From previous studies, it is evident that neural network-based node embedding models are highly scalable to real-world large information networks as well as can be modeled efficiently in unsupervised approach [98, 97]. Therefore, this thesis explores neural network-based unsupervised node embedding models for generating the feature vectors corresponding to nodes.

Let $G(V, E)$ represents the underlying network and $f : V \rightarrow \mathbb{R}^d$ be a mapping function that generates a lower dimensional latent embedding vector for every $v \in V$, then unsupervised network embedding models attempt to solve the following optimization equation [97]:

$$max \sum_{v \in V} logP(\mathcal{N}_v | f(v)) \tag{2.15}$$

where $f(v)$ gives the d dimensional embedding vector for node v, and $\mathcal{N}_v$ represents the neighborhood for node v (node sequence associated with node v) of a given truncated size or context window. Equation (2.15) is also referred to as Skip-Gram model proposed in devising popular word2vec method for word embedding in [104]. Further, the probability of any node in the neighborhood sequence $\mathcal{N}_v$ is assumed to be conditionally independent for the given embedding vector of the source node v. Now the probability P in the Equation (2.15) can be written as

$$P(\mathcal{N}_v | f(v)) = \prod_{u \in \mathcal{N}_v} P(u | f(v))$$

Now similar to [97], assuming the symmetry in embedding space for the source and any node in the neighborhood sequence, the above conditional probability can be estimated using softmax function. Furthermore, the softmax function is

parametrized over the scalar product of embedding vectors of source and neighborhood nodes.

$$P(u|f(v)) = \frac{e^{f(u).f(v)}}{\sum_{w \in V} e^{f(w).f(v)}}$$

Now, with the above assumptions we reformulate the Equation (2.15) as follow:

$$max \sum_{v \in V} \left[-logX_v + \sum_{u \in \mathcal{N}_v} f(u).f(v) \right] \tag{2.16}$$

where $X_v = \sum_{w \in V} e^{f(w).f(v)}$.

Majority of the network embedding models exploiting Equation (2.16) differ in two ways: *(i) considering the sampling strategy for generating node sequences* $\mathcal{N}_v$, and *(ii) handling the computational overhead for* X_v. For example, DeepWalk [98] uses random walks for generating neighborhood for the given node whereas VERSE [103] exploits Personalized PageRank [105] capturing vertex-to-vertex similarity. Further, DeepWalk exploits hierarchical softmax while VERSE uses Noise Contrastive Estimation (NCE) to handle the computational overhead of X_v .

2.7 Summary

This chapter first presents a brief overview over heterogeneous information network. Thereafter, we discuss the local similarity measures and spectral graph kernels. As this thesis exploits centrality of nodes while predicting links, we further discuss state-of-the-art centrality measures. Thereafter, we discuss the unsupervised network embedding framework exploiting neural networks.

Chapter 3

Link Prediction in HIN

The previous chapter briefly discusses some of the existing methodologies used in this thesis such as similarity measures, graph kernels, centrality measures, and network embedding. Majority of the state-of-the-art similarity measures and graph kernels do not incorporate heterogeneous characteristics of the information networks. Thus, these methods treat multiple types of nodes and relations equally and fail to capture the rich characteristics of HINs. Therefore, this chapter proposes heterogeneous transformations of various state-of-the-art similarity measures and spectral graph kernels for link prediction in HIN[1].

3.1 Introduction

Link prediction in HIN is an important problem and garnered a considerable focus in recent times [21, 22, 23, 24, 25]. Unlike homogeneous information network, a HIN is more complex by considering multiple types of objects and relations having different semantics. Further, link prediction in HIN must have the following characteristics [13]:

[1]Though this chapter proposes heterogeneous frameworks for state-of-the-art similarity measures and spectral graph kernels, it gives equal importance to all the nodes while link prediction. We use different forms of node importance and exploit the heterogeneous similarity measures for predicting links in Chapter 4. A comparative study between the proposed heterogeneous setup and its homogeneous counterpart is reported in Chapter 4.

- The link predictors should be capable of predicting links of different types to be coherent with the past and future schema of the HIN.
- The link prediction methods should harness the inter-dependencies among different types of links present in the underlying HIN for collective prediction of future links.

Similarity measures capturing local proximity between the underlying pair of nodes have widely been exploited for link prediction in homogeneous information networks [8]. In particular, similarity measures capturing proximity of path-length two (triadic closure) are mostly popular in addressing link prediction problem. One of the simplest similarity measures based on triadic closure is Common Neighbor (CN) [8] that estimates the likelihood of link existence between a pair of nodes by counting the number of common neighbors they have (refer Section 2.3 for details). Other similarity measures such as Jaccard Coefficient (JC) [35], Adamic-Adar (AA) [36], Resource Allocation (RA) [37] also exploit the estimate of CN score in different perspectives while estimating the similarity between pair of nodes. Although these similarity measures help in predicting future links in homogeneous networks, they are not suitable for link prediction in HIN. It is because these similarity measures treat each common neighbor equally. As discussed above, a HIN has multiple types of nodes and not every node has equal importance while capturing similarity between given pair of nodes. For example, in a bibliographic HIN consisting nodes of types Author, Paper, and Venue, common neighbor of type Paper for given two authors captures a co-authorship characteristics whereas a common neighbor of type Venue states that two authors publishing at the same venue. Thus, it is a requirement to capture the different types of node importance while estimating the likelihood of link existence. Therefore, this chapter proposes heterogeneous formulations of the above-mentioned four local similarity measure namely, CN, JC, AA, and RA capable of capturing importance of different types of common neighbor nodes while estimating the similarity.

The local similarity measures mentioned above exploit the explicit structure of the underlying network to estimate the likelihood of future links. However, they

are limited to capture only the proximity of path-length two. In literature, graph kernels have been used to capture proximity of higher path lengths. Some of the popular graph kernels for homogeneous graphs are Path Counting, Exponential, and Neumann. Therefore, this study further proposes a generalized framework for adapting the existing homogeneous graph kernels to heterogeneous information networks for predicting future links.

3.1.1 Contribution

The major contributions of this chapter are:

- Heterogeneous transformations of homogeneous local similarity measures.
- Heterogeneous transformations for spectral graph kernels.

3.2 Proposed Heterogeneous Local Similarity Measures

Given a heterogeneous network G as defined in Definition (2.1), let $n(x)$ denotes the set of neighbor nodes (of any types) of the node x connected using any link type, $n_c^l(x)$ denotes the set of neighbor nodes of type $c \in \mathcal{V}$ of the node x connected using only link type l, and $w_v(z)$ gives the node importance of node z. Then, we adapt the traditional local node proximity-based methods to heterogeneous information networks as formulated below:

1. **Common Neighbor:** The common neighbor score between two nodes x and y in homogeneous network is defined by the number of common nodes directly incident to the nodes x and y i.e., $S_{CN}(x, y) = |n(x) \cap n(y)|$. In heterogeneous network, the common neighbor score between two heterogeneous

nodes x and y is defined as follows.

$$S_{CN}(x,y) = \sum_{c\in\mathcal{V}}\sum_{l\in\mathcal{E}}\sum_{z\in n_c^l(x)\cap n_c^l(y)} \omega_v(z)$$

With this scoring function, homogeneous CN becomes a special case of heterogeneous CN i.e., considering only one type of nodes and one type of links. For a homogeneous setup of node type c and link type l, the homogeneous CN score between homogeneous nodes x and y i.e., $S_c^l(x,y)$ can be obtained by assigning (i) zeros to all the entries of the node weight vectors and link weight vectors except for the nodes of type c and the links of type l, and (ii) 1 for nodes of type c and the links of type l.

2. **Jaccard Coefficient:** For homogeneous network, traditional JC between two nodes x and y is defined by $S_{JC}(x,y) = \frac{|n(x)\cap n(y)|}{|n(x)\cup n(y)|}$. We define JC for heterogeneous network as follows.

$$S_{JC}(x,y) = \sum_{c\in\mathcal{V}}\sum_{l\in\mathcal{E}} \frac{\sum_{z\in n_c^l(x)\cap n_c^l(y)} \omega_v(z)}{\sum_{z\in n_c^l(x)\cup n_c^l(y)} \omega_v(z)}$$

Like CN, homogeneous JC is a special case of this heterogeneous estimate for the same conditions used for CN.

3. **Adamic Adar:** Traditional AA index between two nodes x and y for homogeneous network is defined as $S_{AA}(x,y) = \sum_{z\in n(x)\cap n(y)} \frac{1}{log(|n(z)|)}$. If $d_c^l(z) = \sum_{i\in n_c^l(z)} w_v(i)$ i.e., sum of the node weights of c type neighbor nodes of z connected by link type l. We define the heterogeneous counterpart as follows.

$$S_{AA}(x,y) = \sum_{c\in\mathcal{V}}\sum_{l\in\mathcal{E}}\sum_{z\in n_c^l(x)\cap n_c^l(y)} \frac{w_v(z)}{\log(1+d_c^l(z))}$$

The 1 in the denominator is the regularization parameter to avoid division by zero. Like CN and JC, homogeneous AA estimate is also a special case of the heterogeneous AA.

4. **Resource Allocation:** RA index between two nodes x and y for homogeneous network is defined as $S_{RA}(x,y) = \sum_{z \in n(x) \cap n(y)} \frac{1}{|n(z)|}$. We then define the heterogeneous counterpart as follows.

$$S_{RA}(x,y) = \sum_{c \in \mathcal{V}} \sum_{l \in \mathcal{E}} \sum_{z \in n_c^l(x) \cap n_c^l(y)} \frac{w_v(z)}{d_c^l(z)}$$

Like CN, JC, and AA, homogeneous RA is also a special case of heterogeneous RA.

3.3 Proposed Heterogeneous Spectral Graph Kernels

As discussed in Section 2.4, state-of-the-art spectral graph kernels exploit Eigenvalue decomposition of the matrix (Adjacency or Laplacian) representing the underlying network. Therefore, to incorporate the heterogeneous characteristics using node importance we modify the adjacency matrix by updating edge weights in the following manner:

For the given network $G(V,E)$, let $\mathbf{A}$ be the adjacency matrix representation of $G(V,E)$, then modified adjacency matrix $\mathbf{H}$, can be derived as

$$\mathbf{H_{xy}} = hm(\omega_v(x), \omega_v(y))\mathbf{A_{xy}} \tag{3.1}$$

where $x \in V$, $y \in V$, and $hm(\omega(x), \omega(y))$ is the harmonic mean of importance/weights of the nodes x and y i.e., $\frac{2.\omega_v(x).\omega_v(y)}{\omega_v(x)+\omega_v(y)}$. The eigenvalue decomposition of the modified adjacency matrix $\mathbf{H}$ can be given as $\mathbf{H} = \mathbf{U_H \Lambda_H U_H^T}$, where $\mathbf{U_H}$ is the matrix consisting of eigenvectors and Λ_H is a diagonal matrix consisting of eigenvalues of matrix $\mathbf{H}$.

1. Path Counting (PC): Spectral transformation of PC kernel for homogeneous graph is defined as $PC(\sum_{i=0}^{k} \mathbf{A}^k) = \mathbf{U}F(\sum_{i=0}^{k} \Lambda^k)\mathbf{U}^T$. Similarly, PC for

heterogeneous graph can be defined using the modified adjacency matrix $\mathbf{H}$ in Equation (3.1) as follows:

$$PC(\sum_{i=0}^{k} \mathbf{H}^k) = \mathbf{U_H} F(\sum_{i=0}^{k} \mathbf{\Lambda_H^k}) \mathbf{U_H}^T \tag{3.2}$$

where $\sum_{i=0}^{k} \mathbf{\Lambda_H}^k$ is defined by a function $f(\lambda_H) = \sum_{i=0}^{k} \alpha_i \lambda_H^i$ for $\alpha_i \geq 0$ and $\alpha_1 > \alpha_2 > \cdots \alpha_{k-1} > \alpha_k$.

2. Exponential (EXP): Spectral transformation of Exponential kernel for homogeneous graph is defined as $EXP(e^{\alpha \mathbf{\Lambda}}) = \mathbf{U} F(e^{\alpha \mathbf{\Lambda}}) \mathbf{U}^T$. Now using the modified adjacency matrix $\mathbf{H}$, heterogeneous transformation of EXP kernel can be defined as follows:

$$EXP(e^{\alpha \mathbf{H}}) = \mathbf{U_H} F(e^{\alpha \mathbf{\Lambda_H}}) \mathbf{U_H}^T \tag{3.3}$$

Similar to PC kernel, Equation (3.3) can be realized using function $f(\lambda_H) = e^{(\alpha \lambda_H)}$.

3. Neumann (NEU): For homogeneous graph, spectral transformation of Neumann kernel is defined as $NEU(\mathbf{A}) = [\mathbf{U} F(\mathbf{I} - \alpha \Lambda) \mathbf{U}^T)]^{-1}$ where $\alpha^{-1} > \lambda^1$. For heterogeneous graph, we define the Neumann kernel as follows:

$$NEU(\mathbf{H}) = [\mathbf{U_H} F(\mathbf{I} - \alpha \mathbf{\Lambda_H}) \mathbf{U_H}^T)]^{-1} \tag{3.4}$$

where $\alpha^{-1} > \lambda_H^1$. Similar to PC and EXP kernels, Neumann kernel can also be defined using function $f(\lambda_H) = \frac{1}{1 - \alpha \lambda_H}$ and $\alpha \neq 0$.

3.4 Applications of the Proposed Methods

The proposed link prediction methods are capable of predicting relations not only between different node types but also incorporating importance of different node types. In this study, effectiveness of the proposed methods are investigated on counter-terrorism problems. In particular, we attempt to study four different

TABLE 3.1: GTD used for constructing heterogeneous terrorist attack network

ID	Country	Region	Province	City	Attack type	Target type	Target subtype	Group	Weapon type	Weapon subtype
1	India	South Asia	Assam	Dibrugarh	Bombing	Educational Institution	School/ University	ULFA	Explosives	Grenade
2	India	South Asia	Orissa	Jajpur	Bombing	Transportation	Bridge/ Car Tunnel	CPI-Maoist	Explosive	Land Mine
3	India	South Asia	Assam	Kokrajhar	Facility	Transportation	Train/ Train Tracks	NDFB	Sabotage	Equipment
4	India	South Asia	J&K	Bijbehara	Bombing	Police	Police Patrol	LeT	Explosives	Vehicle
5	Nigeria	Sub-Saharan Africa	Borno	Maiduguri	Facility	Religious Figures	Place of Worship	Boko Haram	Incendiary	Arson/ Fire
6	Afghanistan	South Asia	Kandahar	Kandahar	Bombing	Religious Figures	Place of Worship	Taliban	Explosives	Suicide

types of aspects in counter-terrorism: *(i) predicting an attack from a terrorist organization on a country, (ii) predicting an attack from a terrorist organization on a city, (iii) predicting target types for terrorist organization, and (iv) predicting preferred weapon-types by terrorist organizations.*

Earlier studies for link prediction on counter-terrorism mainly focus on relation prediction between homogeneous entities such as `terrorist to terrorist` [106, 107], `terrorist organization to terrorist organization [108]`, etc. To the best of our knowledge, the proposed methods are the first link prediction methods which can support relation between different node types in the domain of counter-terrorism[2]. In this chapter, we have investigated effectiveness of predicting all the four types of relations mentioned above. Since the nature of the link prediction is inherently heterogeneous and there are no comparable counterparts, this study reports only the performance of the proposed link prediction methods.

3.5 Dataset

This study uses *Global Terrorism Data (GTD)* collected from *National Consortium for the Study of Terrorism and Responses to Terrorism* [109]. Each terrorist attack is defined by approximately hundred number of features. Out of these features, this study considers ten features which can potentially define an attack

[2]From the year 2011 on-wards, link prediction methods for HIN exploiting meta-paths [21, 22, 23, 24, 25] had been proposed. However, finding best meta-path for the HINs having large number of node and relation types is not a trivial problem. On the other hand, terrorist attacks are often defined by large number of entity types related to each other by multiple semantics. Thus, meta-path-based approaches are not suitable for HINs representing terrorist attacks.

TABLE 3.2: Number of Nodes in Different Classes for Training Data, where, TarType: Target Type, TarSubType: Target Subtype, WeapType: Weapon Type, WeapSubType: Weapon Sub Type, and AttType: Attack Type.

Country	Group	City	Region	Province
186	2562	13034	12	1350
TarType	TarSubType	WeapType	WeapSubType	AttType
22	111	12	28	9

event; *country, region, provincial state, city, attack type, target type, target subtype, terrorist organization, weapon type, and weapon subtype.* These ten features represent ten different node classes in the experimental heterogeneous network. Table 3.1 illustrates a sample of experimental dataset extracted from GTD.

Based on Definition (2.1) and the dataset described in Table 3.1, the experimental heterogeneous terrorist attack network is constructed in the following manner :

1. **Node**: The node set V is defined by unique set of feature instances from all the ten node classes discussed above. For example in Table 3.1 all the field values are treated as a node.

2. **Edge Classes**: Since this dataset consists ten classes of nodes, for an undirected network there can be ${}^{10}C_2 = 45$ classes of edges. Unlike earlier studies which consider few types of edges this work considers every possible edge types. This choice is based on the intuition that every relation between the various node classes collectively defines a terrorist attack event.

3. **Edge**: The edge set E is obtained by putting an edge between all the nodes which belong to same terrorist attack event. For example, the GTD network constructed using Table 3.1 will have edges between India and rest of the node class values value available in the first four rows.

We consider the dataset between 1970 to 2009 as the training and 2010 to 2014 as the test dataset. The total number of nodes in training data is 17326 having 239193 number of relations. Further, Table 3.2 presents the number of instances

in each types of node classes considered in the underlying heterogeneous terrorist attack network. The statistics of test data is presented in Table 3.3.

3.6 Experimental Setup

This section describes the experimental setups adopted by this study for link prediction in GTD dataset using above-discussed local similarity measures and spectral graph kernels.

3.6.1 Link Prediction Using Heterogeneous Local Similarity Measures

Traditional similarity measures defined for homogeneous information networks require only the adjacency representation of the underlying network as input. However, the proposed similarity measures in Section 3.2 require node weights showing node importance as input along with the adjacency representation of the given heterogeneous network.

For the above-discussed dataset in Section 3.5, the adjacency representation of training network, i.e. from 1970-2009 is used as an input to the proposed local similarity measures. Although, the proposed heterogeneous transformations of the similarity measures are capable of incorporating different types of node importance, this chapter treats every node equally, i.e. $\boldsymbol{\omega_v}$ is set to $\mathbf{I^T = \{1, 1, 1.....1\}}$. Further, different types of node importance has been considered in chapter 4. To evaluate, we predict the future links for test network, i.e. from 2010 to 2014.

3.6.2 Link Prediction Using Heterogeneous Spectral Graph Kernels

For link prediction using spectral graph kernels we exploit the spectral evolution model discussed in the study [88]. This framework divides the training network into two parts (i.e. source and validation) and learns the kernel's parameter by fitting best curve from source to validation eigenvalues. In the similar way, we divide the given training network into two networks namely, (i) $\boldsymbol{G^s}$ (1970-2004) and (ii) $\boldsymbol{G^v}$ (1970-2009). The traditional graph kernels defined for homogeneous networks use adjacency representation of the network. In this study, we use the modified adjacency matrix $\mathbf{H}$ obtained using Equation (3.1). Using eigenvalue decomposition (Equation (2.5)), we estimate the eigenvalues and eigenvectors of $\boldsymbol{G^s}$ and $\boldsymbol{G^v}$. As a full rank eigenvalue decomposition requires longer time and few of the leading eigenvalues along with corresponding eigenvectors can well approximate the entire matrix, we use low rank eigenvalue decomposition [3].

Suppose, $\boldsymbol{\lambda^s}$ and $\boldsymbol{\lambda^v}$ represent eigenvalues of $\boldsymbol{G^s}$ and $\boldsymbol{G^v}$ networks, then using the spectral graph kernels defined in Section 3.3 over $\boldsymbol{\lambda^s}$ we learn the optimal kernel's parameters by fitting best curve to the $\boldsymbol{\lambda^v}$. We further apply the optimal parameters of the kernels over $\boldsymbol{\lambda^s}$ to predict the eigenvalues in test network $\boldsymbol{\lambda^t}$ and reconstruct the test output matrix as follows:

Suppose $\mathbf{H^v}$ and $\mathbf{H^t}$ represents the adjacency matrix representing networks $\boldsymbol{G^v}$ (from 1970 to 2009) and $\boldsymbol{G^t}$ (from 1970 to 2014), then

$$\mathbf{H^t} = \mathbf{U_H^v \Lambda_H^t U_H^{v\,T}} \tag{3.5}$$

where $\mathbf{U_H^v}$ is the matrix consisting of eigenvectors of $\mathbf{H^v}$ and $\mathbf{\Lambda_H^t}$ is the matrix consisting of predicted eigenvalues for $\mathbf{H^t}$. The edge weights of $\mathbf{H^t}$ gives the likelihood of existing edges (i.e. from 1970 to 2009) and future edges in test network i.e. from 2010 to 2014.

[3]This study considers rank equal to 30 for eigenvalue decomposition.

TABLE 3.3: Different types of Test Edges considered for Evaluating Link Predictors.

Test Data	gp-cn	gp-cty	gp-tar	gp-weap
All Edges	350	1020	761	413
Missing Edges	104	374	200	71

3.6.3 Evaluation

We evaluate the performance of link prediction using Area under ROC curve (AUC) score given by the following formula [110]:

$$\boldsymbol{AUC = \frac{n_1 + 0.5n_2}{n_{ne} * n_e}} \tag{3.6}$$

where $\boldsymbol{n_1}$ = number of times link prediction score for existing test edge (positive edges) is greater than other non-existing test edges (negative edges), $\boldsymbol{n_2}$ = number of times link prediction score for existing test edge is equal to other non-existing test edges, $\boldsymbol{n_{ne}}$ = number of non-existing test edges, and $\boldsymbol{n_e}$ = number of existing test edges. We assess the performance of link predictors on two different sets of test edges. For the first type of test edges (i.e. All Edges), we consider all the edges appeared between years 2010 to 2014. For the second type of test edges (i.e. Missing Edges), we consider only new edges appeared between 2010 to 2014. By considering the above two types of test edges our intuition is to evaluate the link predictors in terms of:

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

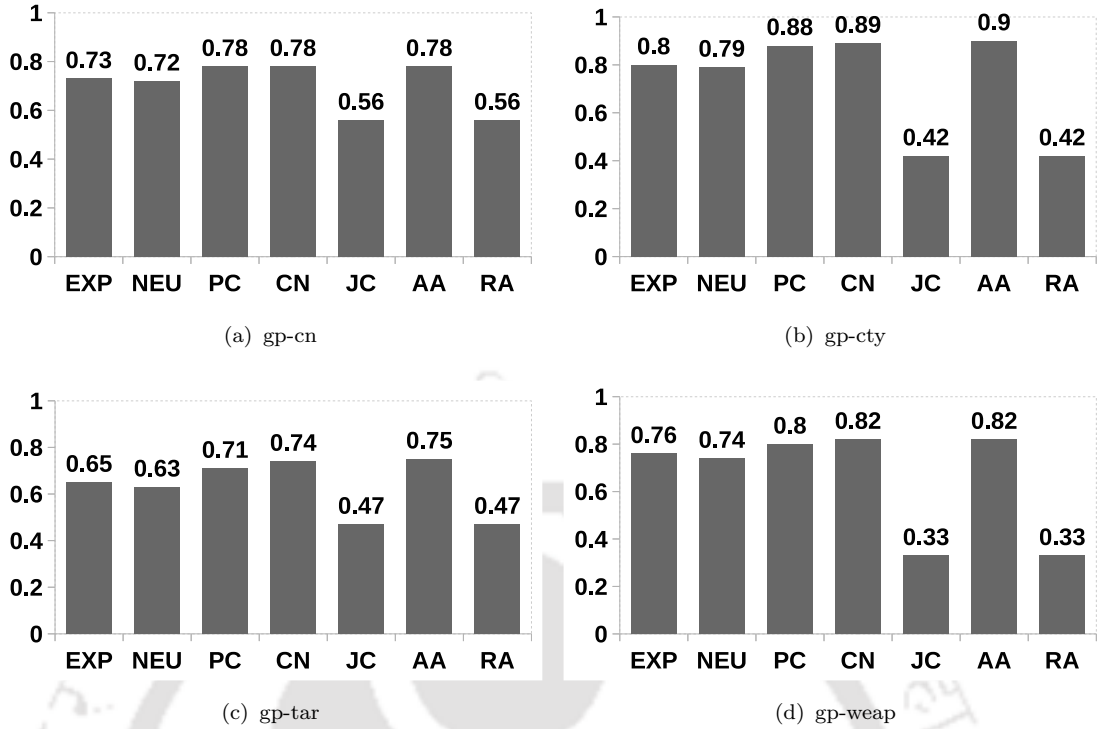
Although the HIN considered in this study have many types of edges, we consider four types of edges for evaluation namely, (i) `gp-cn:` edges between terrorist group to country, (ii) `gp-cty:` edges between terrorist group to city, (iii) `gp-tar:` edges between terrorist group to target type, and (iv) `gp-weap:` edges between terrorist group to weapon type. For all the above four types of existing test edges we generate more than five times of non-existing test edges and estimate the AUC

score. Table 3.3 gives the count of different types of test edges considered under `All Edges` and `Missing Edges`.

3.7 Experimental Observation

In general, the real-world information networks are very sparse and there may be high imbalance in the ratio of existing and non-existing edges. It is observed in the previous studies that evaluation of link prediction by AUC score does not get much affected by the skewed distributions of existing and non-existing edges [111]. Therefore, we exploit AUC score as a metric to assess the performance of all the link predictors. Moreover, an AUC score equals to 0.5 refers to a random predictors. Several studies on link prediction in the domains such as bibliographic networks, social networks, etc. achieve AUC score in the range of 0.65 to 0.9 [112, 37, 113, 111, 114, 24]. Since this is the first study on the prediction of heterogeneous relations in terrorist attack data, we do not have any previous studies to compare as the baselines and we analyze the performance of proposed link predictors on the basis of accepted range of AUC score (i.e. 0.65 to 0.9) by previous studies from different domains.

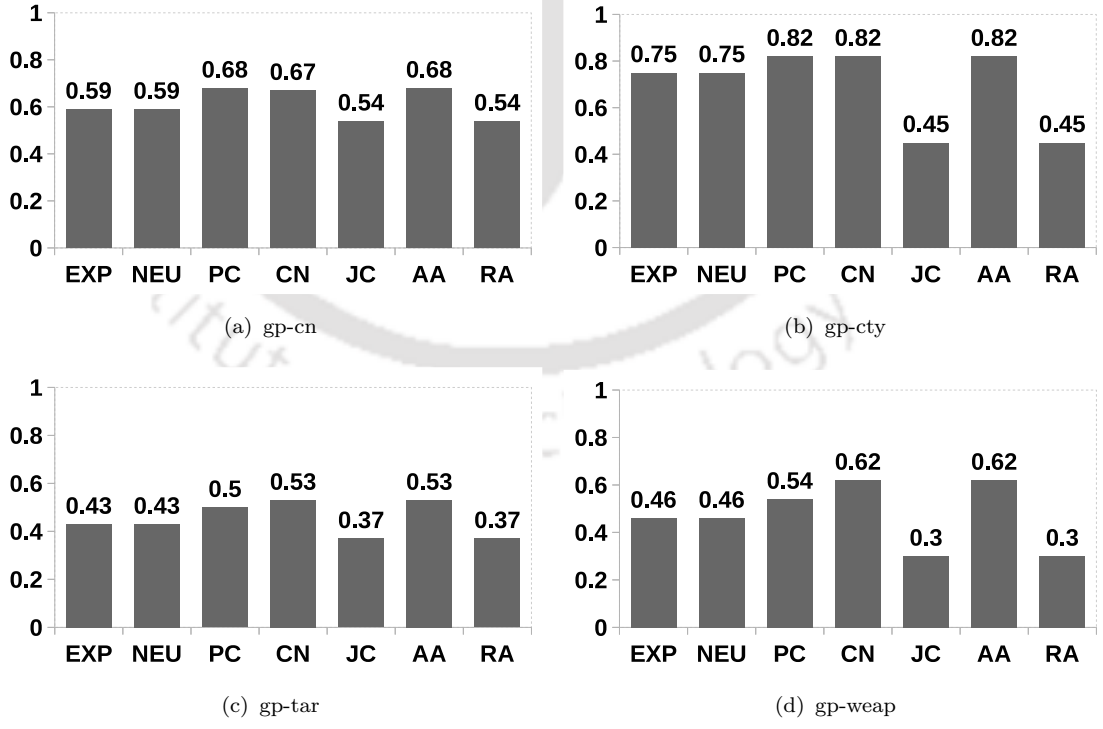
Figures 3.1 and 3.2 present the Average AUC score for all the four types of relations, namely (i) `gp-cn`, (ii) `gp-cty`, (iii) `gp-tar`, and (iv) `gp-weap` for `All Edges` and `Missing Edges` test data respectively. From Figure 3.1, it is evident that all the four types of relations are predicted with convincing AUC score for `All Edges`. For instance, in case of `gp-cn` relation prediction, the best AUC score achieved is 0.78. For predicting `gp-cty` relation, we are able to achieve upto 0.9 AUC score. Further, the best AUC score for `gp-tar` relation is 0.75 whereas in case of `gp-weap` we achieve 0.82 as the best AUC score. From Figure 3.2, it is observed that for link prediction in `Missing Edges`, although relations like `gp-cn` and `gp-cty` achieve the accepted range of AUC score, `gp-tar` and `gp-weap` relation prediction achieve lower AUC values. Further, it is also noted that link predictors achieve lower AUC scores for predicting `Missing Edges` when compared to `All Edges` for all the link



(a) gp-cn (b) gp-cty

(c) gp-tar (d) gp-weap

FIGURE 3.1: Average AUC score for link prediction on All Edges, EXP: Exponential, NEU: Neumann, PC: Path Counting, CN: Common Neighbor, JC: Jaccard Coefficient, AA: Adamic Adar, RA: Resource Allocation.



(a) gp-cn (b) gp-cty

(c) gp-tar (d) gp-weap

FIGURE 3.2: Average AUC score for link prediction on Missing Edges, EXP: Exponential, NEU: Neumann, PC: Path Counting, CN: Common Neighbor, JC: Jaccard Coefficient, AA: Adamic Adar, RA: Resource Allocation.

predictors. This observation is intuitive from the perspective of terrorist attack network as majority of the terrorist attacks are repetitive in nature and follow the past methodologies, target, etc.

For `All Edges`, majority of the link predictors except JC and RA give promising results (refer Figure 3.1). It is convincing to observe that among the good performing link predictors, the least performance achieved in terms of AUC is 0.72 for `gp-cn`, 0.79 for `gp-cty`, 0.63 for `gp-tar`, and 0.74 for `gp-weap`. On the other hand, for `Missing Edges`, all the link predictors except JC and RA achieve the accepted range of AUC score for `gp-cty` relation. Further, in case of `gp-cn`, we observe that three out of seven link predictors give convincing AUC score. One of the reason to have a convincing link prediction performance for `gp-cty` relation (in both types of test edges, i.e. `All edges` and `Missing Edges`) is that the terrorist attack network considered in this study has the highest number of entities from `city` type and second highest from `group` (refer Table 3.2). Thus, it can be inferred that having large repository of terrorist attack data can enhance the performance of predicting future links and may help in proposing counter-terrorism measures.

We now analyze the capabilities of various links predictors used in this study for predicting heterogeneous relations. This study uses two different classes of link predictors, namely (i) local similarity-based and (ii) graph kernel-based. Among the local similarity-based methods, we observe that AA and CN give comparable performance while JC and RA fail to achieve the accepted range of AUC for both types of test data, i.e. `All Edges and Missing Edges`. Among the graph kernels-based methods, we observe that PC performs better than other graph kernels, namely EXP and NEU. Furthermore, Among all the link predictors, AA always gives either comparable or better performance when compared to the rest of the link predictors.

From the above discussed experimental observations, it can be inferred that state-of-the-art local similarity measure and spectral graph kernel-based link predictors

are capable enough to capture the relational dependencies in the underlying heterogeneous terrorist attack network. Furthermore, these link predictors are capable of collective prediction of different types of relations in HIN such as heterogeneous and homogeneous.

3.8 Summary

This chapter focuses on proposing heterogeneous formulations for state-of-the-art local similarity measures and spectral graph kernels. We adapt the traditional common neighbor-based similarity measures such as `Common Neighbor`, `Jaccard Coefficient`, `Adamic-Adar`, and `Resource Allocation` by exploiting the node importance for link prediction in HIN. Further, a heterogeneous framework based on modified adjacency matrix is proposed to adapt the homogeneous spectral graph kernels such as `Path Counting, Exponential,` and `Neumann` for predicting links in HIN. As different node types may have different importance, the proposed heterogeneous versions of similarity measures capture the natural scenarios of the real-world system. This work has been published in SocialCom-2015[4].

In the next chapter, we exploit the above proposed heterogeneous similarity measures for link prediction in a heterogeneous terrorist attack network.

[4]Anil, A., Kumar, D., Sharma, S., Singha, R., Sarmah, R., Bhattacharya, N. and Singh, S.R., 2015, December. "***Link prediction using social network analysis over heterogeneous terrorist network***". In 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity) (pp. 267-272). IEEE, Chengdu, China

Chapter 4

Link Prediction Using Exogenous Factors as Node Importance

The previous chapter proposes heterogeneous versions of local similarity measures and spectral graph kernels for predicting links in a HIN. From the experimental results, it is observed that local similarity measures are simple yet give promising performance. Therefore, this chapter exploits the proposed heterogeneous local similarity measures for link prediction in heterogeneous terrorist attack network. The proposed heterogeneous local similarity-based link predictors harness different types of node importance for estimating the likelihood of a link. Thus, this study attempts to incorporate different types of node importance by exploiting endogenous (e.g. network structure) as well as exogenous (e.g. news media) factors.

Like chapter 3, this chapter also explores terrorist attack dataset. Previous studies on counter-terrorism states that *a terror attack may be influenced by exogenous information such as news media, television, etc. [115, 116].* Thus, in this study, we capture the effects of exogenous factors, namely news reporting and temporal dynamics of the network while predicting future links in the underlying terrorist attack network.

4.1 Introduction

In real-world HINs, a structural change (e.g., formation/deletion of links) might be caused due to endogenous information (existing network structure) or exogenous information (e.g., news media, social media, etc.) [117]. For example, in case of bibliographic HIN, apart from the network structure, a co-authorship may happen between two authors if they follow similar research profiles on twitter[1] or they are featured in the same debate as research experts. Thus, capturing effects of exogenous factors while mining a HIN is an important pre-requisite. Motivated with this, in this chapter we focus on capturing various exogenous information while predicting future links in a Heterogeneous Terrorist Attack Network (HTAN).

Although it is well established that exogenous information affect the evolution of majority of the real world social and information networks [117], yet capturing different types of exogenous factors is a non-trivial research problem. Thus, to incorporate the effect of exogenous information, we propose to use Personalize PageRank (PPR) [105, 118] as a single model capable of estimating the node centrality-based on supervised parametric setups as personalization parameter. In this study, we consider reporting of news media and temporal dynamics of the underlying HTAN as exogenous information and exploit PPR to estimate the node importance personalized to these exogenous information.

Several link prediction methods have been proposed in the literature, e.g. proximity-based [8], path-based [8], tie weight-based [113], random walk-based [119] etc. Except for the few random walk-based approaches, a majority of the methods do not consider node importance while predicting future relations. However, in the context of heterogeneous network (terrorist attack network in particular), node importance may play an important role. For instance, *a terrorist organization may prefer to attack a country which is popular across the globe to achieve more attention*. The heterogeneous similarity measures proposed in Section 3.2 are capable of capturing node importance while estimating the likelihood of a future links. Therefore, this study exploits these heterogeneous similarity measures with

[1]https://twitter.com

various types of node importance capturing endogenous as well as exogenous information for link prediction in HTAN.

4.1.1 Contribution

This chapter has the following major contributions:

- Projecting personalized PageRank as a potential solution to capture various heterogeneous scenarios (without changing underlying model) while estimating network centrality.
- Projecting personalized PageRank as a potential framework to seamlessly incorporate information from exogenous sources.
- Predicting relationship between heterogeneous attributes of the underlying heterogeneous terrorist attack network using heterogeneous similarity measures capable of incorporating exogenous information as node importance.

4.2 Proposed Framework

4.2.1 Node Importance and Centrality Measures

Several centrality measures have been proposed in literature e.g. degree [28], betweenness [29], closeness [30], PageRank [9], Eigenvector [31] etc. As this study aims at incorporating the effects of exogenous information, we consider *Personalized PageRank* (PPR) as a method capable of incorporating exogenous information as its personalization parameter.

4.2.2 Random Walk-Based Method

Given a Markov transition matrix $\mathbf{M}$ of a network $\boldsymbol{G}$, let $\mathbf{x}^{\boldsymbol{t}}$ be a vector such that $\mathbf{x}^{\boldsymbol{t}}[\boldsymbol{i}]$ denotes the probability that a random walker is at the node $\boldsymbol{i} \in \boldsymbol{V}$ after

time $\boldsymbol{t}$. Now, $\mathbf{x}^{\boldsymbol{t}}$ can be defined as below [120].

$$\mathbf{x}^{\boldsymbol{t}}[\boldsymbol{i}] = \sum_{\boldsymbol{j} \in \boldsymbol{V}} \mathbf{x}^{\boldsymbol{t-1}}[\boldsymbol{j}]\mathbf{M}[\boldsymbol{j}, \boldsymbol{i}]$$

Considering for all the vertices in the network, the above expression can be written as

$$\mathbf{x}^{\boldsymbol{t}} = \mathbf{M}\mathbf{x}^{\boldsymbol{t-1}}$$

If the random walker infinitely keeps walking, it may converge to final stationary distribution vector $\mathbf{x}$ i.e.,

$$\mathbf{x} = \mathbf{M}\mathbf{x}$$

If the matrix $\mathbf{M}$ is stochastic and primitive, $\mathbf{x}$ converges to principle eigenvector whose principle eigenvalue is $\mathbf{1}$. Now, a real network such as Web hyperlink, social network, biological network may not be primitive. Brin and Page in their seminal paper [9] address this problem by introducing a damping factor $\boldsymbol{d}$ that allows the random walker to stop the walk and restarts from a random node, and propose the popular *PageRank* centrality measure. Thus, the PageRank of a node $\boldsymbol{i} \in \boldsymbol{V}$ is defined as below.

$$\mathbf{x}[\boldsymbol{i}] = \boldsymbol{d} \sum_{\boldsymbol{j} \in \boldsymbol{V}} \mathbf{x}[\boldsymbol{j}]\mathbf{M}[\boldsymbol{j}, \boldsymbol{i}] + (\mathbf{1} - \boldsymbol{d})\frac{\mathbf{1}}{\boldsymbol{N}}$$

where $\boldsymbol{N}$ is the number of nodes in the network and the damping factor $\boldsymbol{d}$ is the probability of continuing random walk by following the links in the network and $(\mathbf{1} - \boldsymbol{d})$ is the probability of jumping to a random node. In other words, we can write it as below.

$$\mathbf{x} = [\boldsymbol{d}\mathbf{M} + (\mathbf{1} - \boldsymbol{d})\mathbf{I}\mathbf{v}^{\boldsymbol{T}}]\mathbf{x} \tag{4.1}$$

where $\mathbf{I}$ is the identity vector i.e., $\mathbf{I}^{\boldsymbol{T}} = \{\mathbf{1}, \mathbf{1}, \mathbf{1}.....\mathbf{1}\}$ and $\mathbf{v}^{\boldsymbol{T}} = \{\frac{\mathbf{1}}{\boldsymbol{N}}, \frac{\mathbf{1}}{\boldsymbol{N}}, ..., \frac{\mathbf{1}}{\boldsymbol{N}}\}$. By adding the second term, the transition matrix $\boldsymbol{d}\mathbf{M} + (\mathbf{1} - \boldsymbol{d})\mathbf{I}\mathbf{v}^{\boldsymbol{T}}$ is primitive.

In the above PageRank formulation, a random walker jumps to any random node with equal probability i.e., $\frac{\mathbf{1}}{\boldsymbol{N}}$. However, in a heterogeneous network, different classes of nodes may have different privileges. We may wish to determine centrality

of a network or similarity between nodes by considering certain classes of the node with higher importance. Thus, we have adopted *Personalised PageRank* to address the problem of centrality in a heterogeneous network.

4.2.3 Personalised PageRank for heterogeneous network

Unlike PageRank, personalised PageRank assumes that different user will have a different walking preference. A person in the film industry may prefer to visit web pages in entertainment domain more than literature. In such a scenario, traditional PageRank fails to capture user's preferences. Personalized PageRank, as discussed in [118], incorporates user's preferences and can be defined as follows.

$$\mathbf{x} = [d\mathbf{M} + (1-d)\mathbf{I}\mathbf{p}^T]\mathbf{x} \tag{4.2}$$

where $\mathbf{p}$ is a stochastic personalized vector reflecting user's preference of visiting the nodes in random jump i.e., $\mathbf{p}[i]$ is the probability of jumping to node i by the user.

Let $\mathbf{B}$ be the weighted adjacency matrix of the heterogeneous graph $\boldsymbol{G}$. In general, matrix $\mathbf{B}$ may not be stochastic. We then perform the following transformation to make it a stochastic transition matrix.

$$\mathbf{B}[i,j] = \begin{cases} \dfrac{\omega_e(i,j)}{\sum\limits_{k \in V} \omega_e(i,k)}, & \text{if } \sum\limits_{k \in V} \omega_e(i,k) \neq 0 \\ \dfrac{1}{|V|}, & \text{Otherwise} \end{cases}$$

Now, let $\mathbf{q}$ be a personalized stochastic vector representing the node priorities (or importance) as defined by $\boldsymbol{\omega_v}$ of $\boldsymbol{G}$ i.e., $\mathbf{q}[i] = \dfrac{1 + \boldsymbol{\omega_v}[i]}{\sum_{k \in V} \boldsymbol{\omega_v}[k] + |V|}$. We provide smoothing parameter to avoid zero probability. Thus, we formulate personalized PageRank centrality for a heterogeneous graph by substituting $\mathbf{M}$ by matrix $\mathbf{B}$ and $\mathbf{p}$ by $\mathbf{q}$ in Equation (4.2).

$$\mathbf{x} = [d\mathbf{B} + (1-d)\mathbf{I}\mathbf{q}^T]\mathbf{x} \tag{4.3}$$

Proposition 4.1. *The matrix* $d\mathbf{B}+(1-d)\mathbf{I}\mathbf{q}^T$ *in the Equation (4.3) is stochastic and primitive.*

Proof. Since $\mathbf{q}$ is a stochastic matrix with non-zero elements, the matrix $\mathbf{I}\mathbf{q}^T$ is a stochastic matrix with non-zero elements. It has no zero elements i.e., it represents a completely connected graph. Therefore, $\mathbf{I}\mathbf{q}^T$ is irreducible and aperiodic. Further, the matrix $\mathbf{B}$ is stochastic by its construct. Since convex summation of two stochastic matrices is a stochastic matrix, the matrix $d\mathbf{B} + (1 - d)\mathbf{I}\mathbf{q}^T$ is a stochastic matrix. Further, since $\mathbf{I}\mathbf{q}^T$ is irreducible and aperiodic, the convex sum is also a irreducible and aperiodic matrix. Hence, the matrix $d\mathbf{B} + (1 - d)\mathbf{I}\mathbf{q}^T$ is stochastic and primitive. □

Without loss of generality, we can realize that the centrality vector $\mathbf{x}$ in Equation (4.3) will converge to unique principle eigenvector of $d\mathbf{B} + (1 - d)\mathbf{I}\mathbf{q}^T$ whose eigenvalue is $\mathbf{1}$. We use the centrality score of Equation (4.3) to define estimate of $w_v(x)$ for the node proximity based link prediction.

4.2.4 Incorporating Exogenous information

As mentioned in the earlier discussion, apart from the topological information, we extend our study with other information such as appearance in news publications, temporal characteristics of network entities. Such information may be available in terms of a vector or a matrix. The proposed PPR has the capability to incorporate such information without any problem. If the information is available in the form of a node vector $\mathbf{r}$, Equation (4.3) can be extended as

$$\mathbf{x} = [d_1\mathbf{B} + d_2\mathbf{I}\mathbf{q}^T + d_3\mathbf{I}\mathbf{r}^T]\mathbf{x} \tag{4.4}$$

where $d_1 + d_2 + d_3 = 1$. Similarly, if the information is available in the form of a matrix $\mathbf{R}$, Equation (4.3) can be extended as

$$\mathbf{x} = [d_1\mathbf{B} + d_2\mathbf{I}\mathbf{q}^T + d_3\mathbf{R}]\mathbf{x} \tag{4.5}$$

TABLE 4.1: Characteristics of dataset and exogenous information.

		Node	Edge	Duration
GTD	**Training**	7177	92858	2000-2009
	Test(gp-cn)	258	231	2010-2014
	Test(gp-cty)	758	740	2010-2014
	Test(gp-tar)	179	522	2010-2014
THE HINDU	**Training**	7177	9699	2000-2009
Tensor GTD matrix	**Training**	7177	92858	2000-2009
Tensor activeness vector	**Training**	7177	NA	2000-2009

TABLE 4.2: Number of Nodes in Different Classes for Training Data, TarType: Target Type, TarSubtype: Target Subtype, WeapType: Weapon Type, Weap-Subtype: Weapon Sub Type, and AttType: Attack Type.

Country	**Group**	**City**	**Region**	**Province**
126	707	4751	12	1411
TarType	**TarSubtype**	**WeapType**	**WeapSubtype**	**AttType**
20	106	9	27	8

Thus, the expression can be further extended if additional information is available from multiple sources.

4.3 Dataset and Experimental Setup

This study uses *Global Terrorist Data (GTD)* collected from *National Consortium for the Study of Terrorism and Responses to Terrorism* [109]. The heterogeneous terrorist attack network (HTAN) is constructed by following the similar procedure adopted in Chapter 3.

For evaluation, we have considered the dataset between 2000 to 2009 as the training and 2010 to 2014 as the test dataset. Table 4.1 shows the characteristics of the datasets. Table 4.2 provides the count for different types of nodes used for training in this study.

As discussed in Section 4.1, exogenous information affect the underlying network structure. Thus, Co-occurrence of different objects defining a terrorist attack in news articles justifies the existence of relational dependencies among them.

Therefore, to investigate the effect of news media this work exploits the published news articles in popular newspaper from India namely, THE HINDU[2]. We generate a new network which only considers nodes and edges from GTD network. If two nodes are co-occurring in an article we put an edge between the node pair. Since this network is built particularly over the GTD network with THE HINDU, we name this network as $\boldsymbol{GTD_{HINDU}}$.

From earlier studies [121, 122], it can be inferred that incorporating temporal dynamics of the underlying network may result in better performance in estimating centrality, link prediction, etc. For example in counter-terrorism, recent activity of terrorist organizations may be of more importance than their past behavior. To incorporate this we further create a tensor GTD network ($\boldsymbol{GTD^T_{Freq}}$) and a Tensor activeness vector ($\boldsymbol{GTD^T_{Act}}$) which give more importance to recent relationships in GTD and node frequency from GTD respectively. For constructing $\boldsymbol{GTD^T_{Freq}}$, we divide the GTD network into ten different snapshots where a particular snapshot represents a network for one year (i.e. 2000, 2001, etc.). In the similar way, for $\boldsymbol{GTD^T_{Act}}$, we divide the frequency of different nodes year wise which results in ten activeness vectors. Now, $\boldsymbol{GTD^T_{Freq}}$ and $\boldsymbol{GTD^T_{Act}}$ are constructed in the following manner:

Suppose the dataset is divided in $\mathbf{X_1}$ to $\mathbf{X_T}$ matrices, then the collapsed data from 1 to T is:

$$\mathbf{E} = \sum_{\mathbf{t=1}}^{\mathbf{T}} \boldsymbol{\alpha}^{\mathbf{(T+1)-t}} \mathbf{X_t} \tag{4.6}$$

where $\boldsymbol{\alpha} \in \mathbf{(0, 1)}$ can be chosen by the user. In this study, we have considered 0.8 as the value of $\boldsymbol{\alpha}$. Similarly, we create the tensor vector by considering E and X as vectors.

[2]https://www.thehindu.com/ (We consider THE HINDU in this study as it has consistent archived news reporting related to the world-wide terrorist attack for the duration being studied, i.e., 2000-2009)

4.3.1 Parameters Used for Ranking Nodes Using PPR

As discussed in Section 4.1, the objective of this work is to study the effects of exogenous information such as, news media ($\boldsymbol{GTD_{HINDU}}$) and temporal dynamics of GTD network ($\boldsymbol{GTD^T_{Freq}}$, $\boldsymbol{GTD^T_{Act}}$). To incorporate the influence of exogenous information this study proposes to use node importance obtained using PPR personalized by various exogenous information. From PPR formulation in Equation (4.3), it is evident that it can be adapted for both types of personalization parameter, i.e. matrix and vector by just replacing $\mathbf{Iq}^T$ to $\boldsymbol{GTD_{HINDU}}$ or $\boldsymbol{GTD^T_{Freq}}$ matrix and $\mathbf{q}$ to $\boldsymbol{GTD^T_{Act}}$ vector respectively. Utilizing these three external information and two more vectors consisting of total frequency of nodes (from 2000 to 2009, refer Table 4.1) and only the recent node frequency (i.e. from 2009) from GTD network, we propose the following five parametric variants of PPR centrality:

1. $\text{PPR}^{\boldsymbol{Co}}$: PPR with co-occurrence matrix obtained using $\boldsymbol{GTD_{HINDU}}$ as personalization parameter.
2. $\text{PPR}^{\boldsymbol{T}}_{\boldsymbol{Rel}}$: PPR with $\boldsymbol{GTD^T_{Freq}}$ matrix as personalization parameter.
3. $\text{PPR}^{\boldsymbol{T}}_{\boldsymbol{Act}}$: PPR with $\boldsymbol{GTD^T_{Act}}$ vector as personalization parameter.
4. $\text{PPR}_{\boldsymbol{Freq}}$: PPR with frequency of nodes during the training period, i.e., during 2000-2009 as personalization vector.
5. $\text{PPR}_{\boldsymbol{Recy}}$: PPR with the frequency of nodes in recent past is considered as personalization vector. To estimate the recent frequency of nodes, we select the total frequency of nodes in last one year from training data, i.e. 2009. We chose year 2009 based on the intuition that, node frequency in the nearest year from test data would provide more correct information. However, recent information may also be fine-grained by treating node frequency differently for different days and months using some weighting parameters. For example, while estimating total frequency for the year 2009 one may weight more to the frequency in specific months, nearby dates from a past attack, etc.

4.4 Experimental Observations

This section investigates the ability of parametric Personalized PageRank formulated in Section 4.2.3 and Section 4.2.4 for performing various network-mining tasks without changing the underlying model, but by passing appropriate user-defined personalization parameters. We begin with investigating effect of different personalization parameters of PPR on node importance. Here we analyze the linear dependency between the rank obtained by the model for terrorist organizations with their actual activeness frequency in future (from 2010 to 2014) using Pearson Correlation Coefficient. We further attempt to understand the effect of node importance exploiting exogenous information on link prediction using heterogeneous similarity measures defined in Chapter 3 (refer Section 3.2). The link prediction analysis is focused mainly on three heterogeneous relations; (i) relation between terrorist organization and country indicating a future possible attack by organization on a country, `gp-cn`, (ii) relation between terrorist organization and city indicating a future possible attack by an organization on city, `gp-cty`, and (iii) relation between terrorist organization and a target type in future attacks, `gp-tar`. Further, we also investigate the effect of proposed models for predicting relation between homogeneous attributes, namely, `gp-gp` which shows the adaptability strength of the models.

4.4.1 Centrality and its correlation with future activities

In network science, centrality measures define the level of prominence of a node in a network. Among various centrality measures, PPR is a flexible random walk-based model where the walker visits the nodes under a supervised direction. This section investigates the effect of various formulations of PPR model parameters which potentially provide high correlation with the ground truth (i.e., future activities). For this task, we consider all the active terrorist organizations during testing period (i.e., between 2010 to 2014) and rank them by their frequency of attacks. This ordered list is considered as the ground truth. We, then, estimate

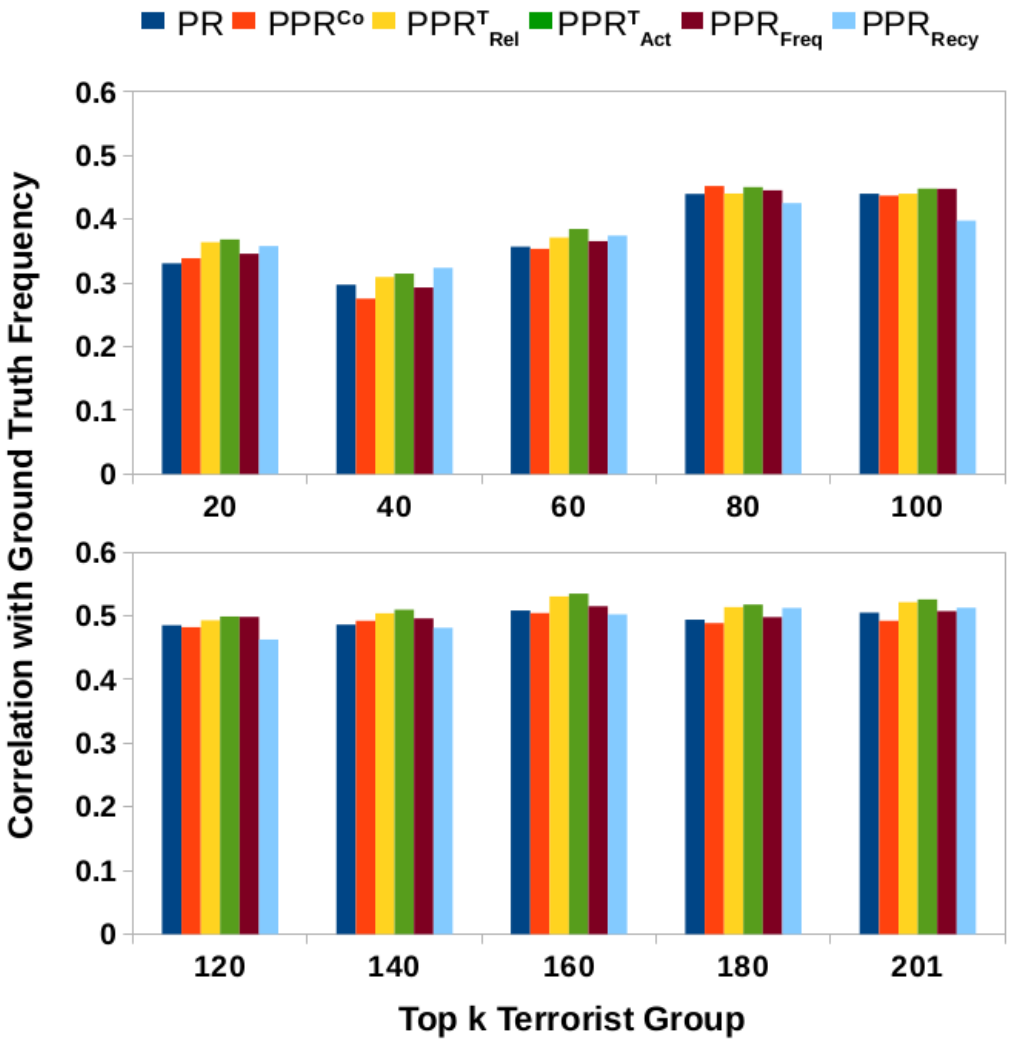


FIGURE 4.1: Correlation of different ranking models with future attack frequencies (Ground Truth), $K \in (20, 40, 60, ..., 201)$.

Pearson's correlation between the ground truth and the observed PPR centrality order (estimated from the training data) using various personalized parameters. We investigate five PPR model parameters as described in Section 4.3.1. These five model parameters are formulated to study the effect of four different aspects; (i) effect of media reporting ($\boldsymbol{PPR^{co}}$), (ii) effect of recent activities ($\boldsymbol{PPR_{Recy}}$), (iii) effect of frequency of activities in the past ($\boldsymbol{PPR_{Freq}}$), and (iv) temporal effect ($\boldsymbol{PPR^{T}_{Rel}}$, $\boldsymbol{PPR^{T}_{Act}}$). Apart from these PPR variants, we also investigate correlation performance of non-personalized PageRank model.

The plots in Figure 4.1 show correlation between ranking of terrorist organizations using training data by PPR and ranking by the frequency of attacks during testing

TABLE 4.3: Link Prediction performance in Average AUC score by different parametric PPR models and state-of-the-art methods (do not use node importance, Unweighted).

Similarity Measure	Relation	Weighted/Ranking Models						
		PR	**PPR^{Co}**	**PPR$^{T}_{Rel}$**	**PPR$^{T}_{Act}$**	**PPR$_{Freq}$**	**PPR$_{Recy}$**	**Unweighted**
Common Neighbor	gp-cn	0.741	0.739	0.74	0.741	0.736	**0.75**	0.74
	gp-cty	0.702	0.683	0.71	0.711	0.703	**0.732**	0.712
	gp-tar	0.73	0.712	0.726	0.727	0.725	0.732	**0.747**
Jaccard Coefficient	gp-cn	0.649	0.649	0.679	0.68	0.679	**0.683**	0.561
	gp-cty	0.547	0.569	0.581	0.581	0.58	**0.588**	0.436
	gp-tar	0.62	0.65	0.675	0.677	0.672	**0.689**	0.612
Adamic Adar	gp-cn	0.743	0.738	0.741	0.742	0.737	**0.75**	0.736
	gp-cty	0.705	0.682	0.712	0.713	0.703	**0.733**	0.695
	gp-tar	0.733	0.715	0.727	0.728	0.725	0.733	**0.741**
Resource Allocation	gp-cn	0.743	0.737	0.743	0.744	0.74	**0.753**	0.7
	gp-cty	0.711	0.68	0.715	0.717	0.708	**0.74**	0.52
	gp-tar	**0.739**	0.719	0.729	0.73	0.728	0.737	0.687

period. The motivation for the correlation analysis is to assess the linear dependency between the ranking by proposed models and actual activities by terrorist organizations in future. It is evident from Figure 4.1 that appropriate personalization may enhance correlation performance. Moreover, it is observed that over all ranges of the top organization, at least three parametrized models outperform non-parametrized model.

From Figure 4.1, it is observed that temporal-based parameters ($\boldsymbol{PPR^{T}_{Rel}}$ and $\boldsymbol{PPR^{T}_{Act}}$) outperform other models. It also shows that future activities are influenced by their past activities prioritized by their recent activities (tensor formation in Equation (4.6) giving higher importance to recent activities.).

Remarks: Topological information of a network alone is not sufficient to predict future activities of a terrorist organization. Considering additional information such as historical activities of the organization, media appearance and recent activities etc. along with topological structure help in predicting future activities.

4.4.2 Heterogeneous Relationship Prediction

In the above section, we have investigated correlation between future activities of an organization and topological centralities, historical activities, recent activities,

and analyze their responses in predicting future activities. With the close inspection of GTD, we found that terrorist organizations mostly follow their historical trends, i.e. attacking same country, adopting similar target types, etc. Thus, we predict repeated future relationships between a terrorist organization and other objects defining an attack. The HTAN considered in this study is undirected and weighted having ten types of nodes and forty five types of relations.

The proposed link prediction frameworks in Section 3.2 can potentially study link prediction between all forty five relations. However, this study focuses on three types of relationships; (i) organization attacking a country (`gp-cn`) consisting 231 relations between 201 terrorist groups and 57 countries, (ii) organization attacking a city (`gp-cty`) consisting 740 relations between 102 terrorist groups and 656 cities, and (iii) organization attacking on a target type (`gp-tar`) consisting 522 relations between 169 terrorist group and 10 target types. We consider all the four types of heterogeneous similarity measures defined in Section 3.2; Common Neighbor (CN), Jaccard Coefficient (JC), Adamic Adar (AA) and Resource Allocation (RA) and predict the above mentioned three types of relationships (`gp-cn, gp-cty, gp-tar`). Considering the above five PPR-based measure of centrality as well as non parametrized PageRank as node weights, we investigate their link prediction performance using average AUC score.

Table 4.3 presents the link prediction performances in terms of Average AUC score for six node centrality setups with the baseline prediction models without considering node weights (Unweighted). Out of 72 prediction cases (six centralities, four prediction models, and three types of relations to be predicted), the proposed node weighed version outperforms unweighted counterparts at least in 70% cases. It shows that node weight helps in improving prediction accuracy. Among the four prediction models, RA dominates the other three models (CN, JC, and AA) in all the parametric setups except for PPR^{Co}. While comparing the performance of node weighted RA with that of unweighted counterpart, it clearly shows that node weighted RA (with all six centralities) outperforms the unweighted counterpart. From Table 4.3, it is evident that node importance has significant contribution

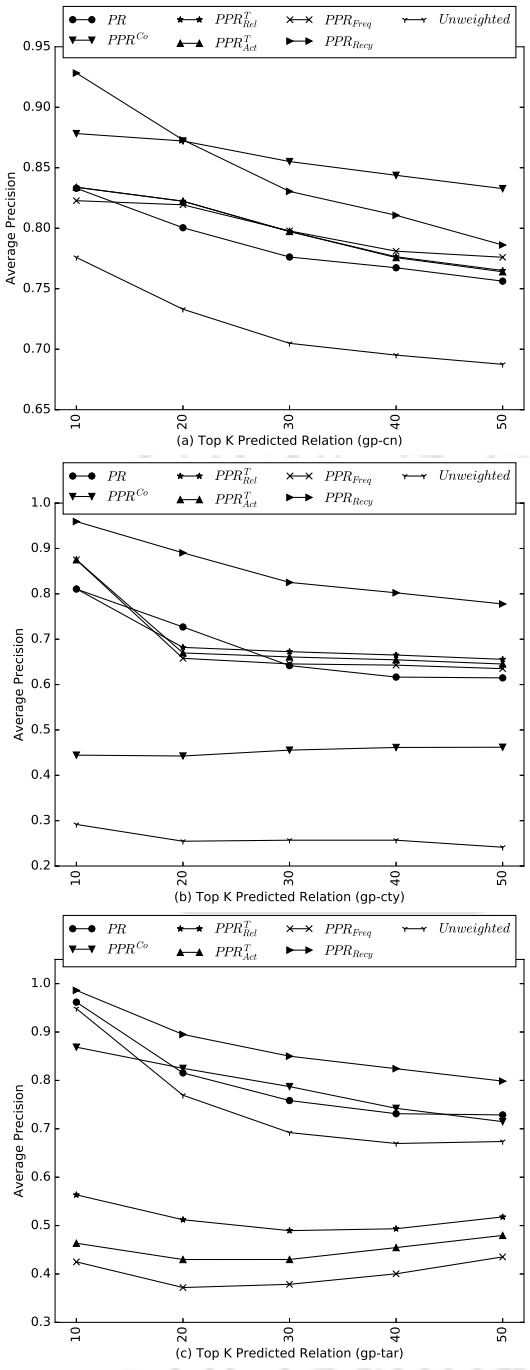


FIGURE 4.2: Average Precision for top 50 predicted relations by RA.

while predicting future repeated relations. Prediction score for RA is improved by 7.5% for `gp-cn`, 42.4% for `gp-cty` and 7.6% for `gp-tar`.

While observing significant improvement in prediction performance after considering node importance (as compared to unweighted counterpart), it is observed that the six personalization parameters provide comparable AUC scores. Though

we observe significant differences in correlation analysis as reported in Figure 4.1, the comparable prediction performance in Table 4.3 may be due to the averaging effect while estimating AUC scores. To investigate this, we further evaluate the predicted relations using Average Precision at K ($\boldsymbol{AP@K}$). $\boldsymbol{AP@K}$ estimate prediction performance for some of the top predictions (in terms of likelihood of link existence) and can be defined as follows:

$$\boldsymbol{AP@K} = \frac{\sum_{k \in \mathcal{P}} \boldsymbol{Precision@k}}{|\mathcal{P}|} \tag{4.7}$$

where $\mathcal{P}$ is set of positions of the correct predictions in top $\boldsymbol{K}$.

Among the four link prediction methods we chose RA as it is atleast 70% times better than others (see Table 4.3).

Figure 4.2 shows the AP@$\boldsymbol{K}$ for different values of $\boldsymbol{K}$, i.e. 10, 20, 30, 40, and 50. It is evident that weighted methods (PR, PPR) perform better than unweighted method in predicting all the three relations, i.e. `gp-cn`, `gp-cty`, `gp-tar` in Figure 4.2 (a), (b), and (c) respectively. It can be inferred from this observation that with appropriate prioritization PageRank-based methods can be exploited for better predictions. In almost all the cases $\text{PPR}_{\boldsymbol{Recy}}$ performs better than others which shows the ability of recent information in predicting future relations. It is also observed that for predicting `gp-cn` relation $\text{PPR}^{\boldsymbol{Co}}$ performs better than all others. This observation indicates that co-occurrence of terrorist organization and country on news media influence future attacks. However, $\text{PPR}^{\boldsymbol{Co}}$ is observed to be under-performing for predicting `gp-cty` and `gp-tar` relations. It may be due to less frequency of co-occurrences for city or target types with the terrorist organizations in news articles as the experimental dataset limits to THE HINDU publication only which may have missing information for world-wide terrorist attack information at finer granularity such as city and target types. Thus, a world-wide and diverse media collection from multiple sources such as news media, social media, television, etc. may provide more stronger inference.

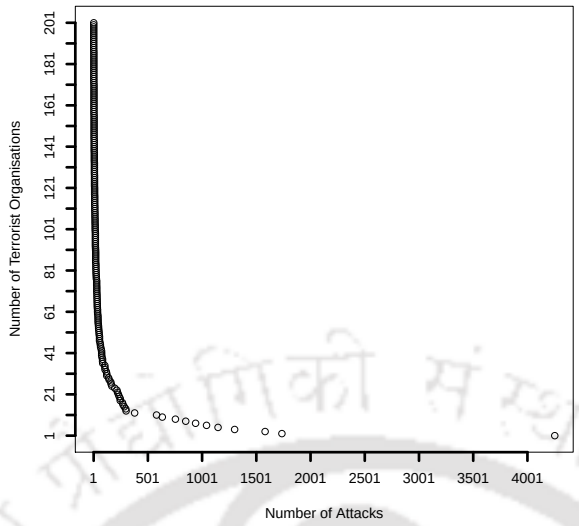


FIGURE 4.3: Terrorist Attack Distribution over 2000-2014.

In the above discussion, we have averaged the scores over the entire set of terrorist organizations. However, attack frequency of different organization varies widely. Figure 4.3 shows attack frequency distribution of different organizations. It clearly shows that very few terrorist organizations carry out majority of the attacks globally. Characteristics of active and less active organizations may be different. To understand such possible patterns, we further analyze $\boldsymbol{AP@K}$ for most active and less active terrorist organizations separately in two ways: *(i) Considering only those relations formed by top ten most active terrorist groups, (ii) Relations formed by other least active terrorist groups.*

Similar to the observation in Figure 4.2, weighted methods perform better than unweighted in almost all the cases for predicting relations contributed by most active and the rest terrorist organizations (refer Figure 4.4). It can also be seen that $\text{PPR}_{\boldsymbol{Recy}}$ constantly performs better than all the setups while considering relations only from top ten most active groups. However, it always under-performs for the relations considering other least active groups. This observation is intuitive as $\text{PPR}_{\boldsymbol{Recy}}$ utilizes the recent frequency of terrorist organizations as personalization vector to PPR. It is interesting to note from Figure 4.4 that, $\text{PPR}^{\boldsymbol{Co}}$ underperforms in predicting relations from most active organizations while it constantly performs better in predicting relations by other least active organizations. Hence,

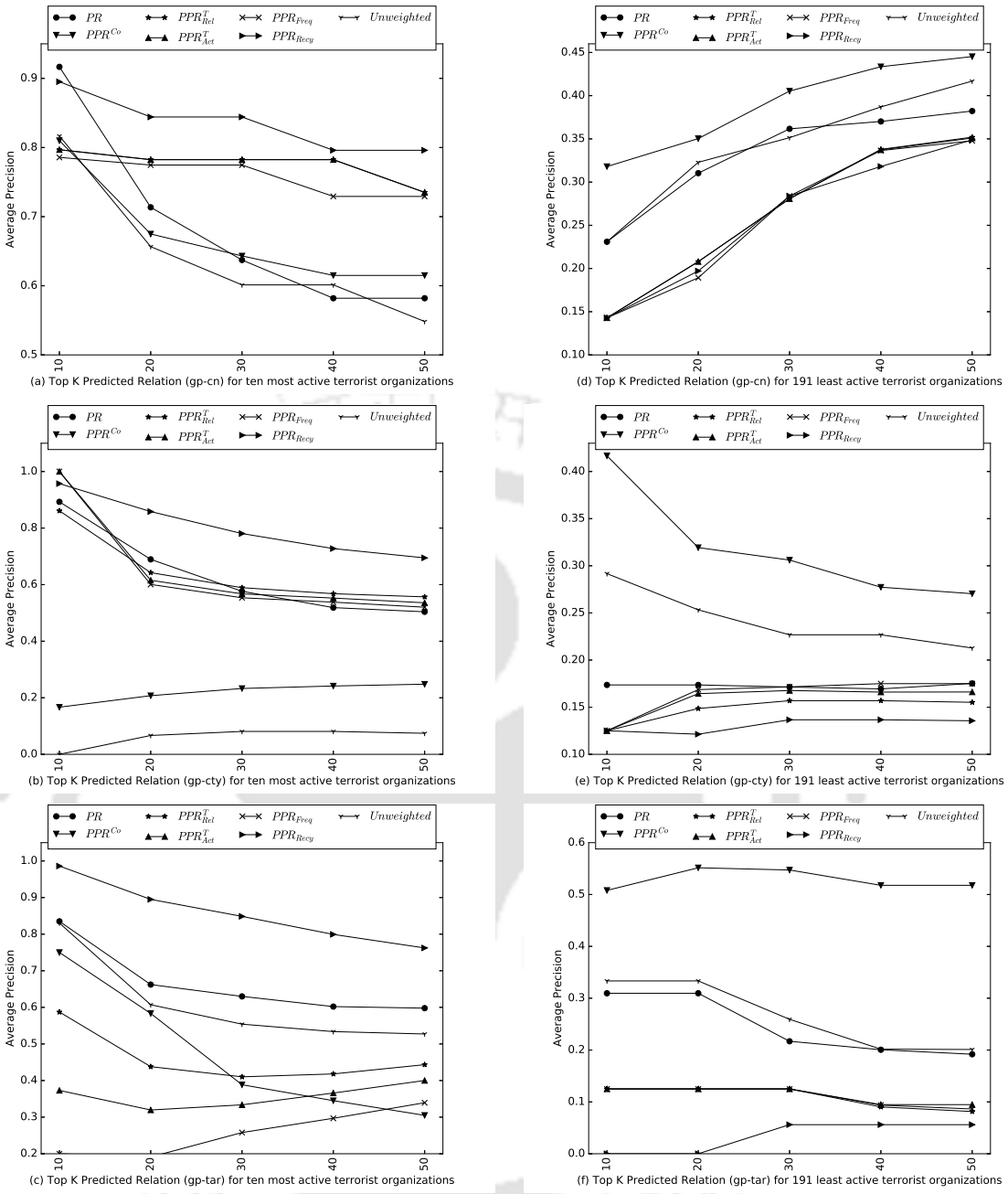


FIGURE 4.4: Average Precision by RA for Top 50 predicted relations formed by Most Active and Rest of the terrorist organizations.

it may be inferred that in case of most active terrorist organizations recent attack information is crucial whereas media reporting helps in predicting relationship formed by least active terrorist organizations.

Remarks: Predicting relationship gets better by considering the effect of exogenous information as node weights. However different information sources have different effects. The recent information is important for most active terrorist

TABLE 4.4: Top five Terrorist Organization Alliance predicted by RA using all the parametric variants of PPR as centrality measures listed in Subsection 4.3.1.

PR		PPR^{Co}		PPR^{T}_{Rel}		PPR^{T}_{Act}		$Freq$		$Recy$		$Unweighted$	
Taliban	TTP	**LTTE**	**Maoists**	Taliban	TTP	Taliban	TTP	Taliban	TTP	Taliban	TTP	**AQLIM**	**GSPC**
Maoists	**CPI-Maoist**	**HM**	**LeT**	**Maoists**	**CPI-Maoist**	**Maoists**	**CPI-Maoist**	**Maoists**	**CPI-Maoist**	**Maoists**	**CPI-Maoist**	**Hamas**	**PIJ**
LTTE	Taliban	**LeT**	**JeM**	LTTE	Taliban	LTTE	Taliban	LTTE	Taliban	LTTE	CPI-Maoist	**ELN**	**FARC**
LTTE	**Maoists**	**HM**	**JeM**	**LTTE**	**Maoists**	**LTTE**	**Maoists**	**LTTE**	**Maoists**	**LTTE**	**Maoists**	FARC	AUC
LTTE	TTP	**LTTE**	**ULFA**	**LTTE**	**CPI-Maoist**	**LTTE**	**CPI-Maoist**	LTTE	TTP	LTTE	Taliban	MILF	NPA

organization while news media helps in tracking down other least active organizations.

4.4.3 Predicting Top Alliance among various Terrorist Organizations

The terrorist attack network and link prediction methods explored in this study can also be utilized for analyzing potential alliances among terrorist organizations without modifying the underlying framework. It is because, two terrorist organizations may be allies if they have common targets or share a common ideology. Table 4.4 lists the top five alliances obtained using RA for different parametric PPR as centrality measures. After referring to external sources; (i) *Wikipedia pages related to concerned terrorist organizations* and (ii) *IDSA publications towards the possible alliance*, we have manually verified the top observed alliances reported in Table 4.4. It is observed that almost 60% of the predicted alliances are found to be true (highlighted in Table 4.4). For example, it is easy to verify from above sources that Maoist, CPI-Maoist, and LTTE help each other in various needs, such as training, finances, etc., and are potential allies[3]. In addition, it is also observed that PPR^{Co} outperforms other prediction models. It shows that incorporating news media reporting helps in determining organizational alliances in a better way.

[3]http://www.idsa.in/idsacomments/Maoistsglobalweboflinkages

4.5 Summary

This chapter investigates the ability of personalized PageRank in incorporating exogenous information as node importance for link prediction in a heterogeneous terrorist attack network. In particular, we study the effects of exogenous information on predicting terrorist organization's future activity and prediction of the relationship between a terrorist organization with other objects (country, city and target types). From various parametric setups, we observe that PPR when personalized with network's temporal dynamics performs better in predicting future attacks of terrorist organizations. Further it is also observed that incorporating node importance can enhance relationship prediction performance. Moreover, it is evident that recent information when used as a personalization parameter improves the relationship prediction. This chapter has been first published in conference proceedings of Web Intelligence[4] and the extended version is published at Web Intelligence Journal[5].

Although, the heterogeneous local similarity measures help in predicting future relations in the considered HIN, they explicitly exploit the network structure. However, in real-world HINs two nodes may be serving the similar roles even though they are not connected. Therefore, in the next chapter, we extend our study to explore network embedding frameworks capable of generating node features in latent space.

[4]Anil, A., Singh, S.R. and Sarmah, R., 2016, October. "***Personalised PageRank as a Method of Exploiting Heterogeneous Network for Counter Terrorism and Homeland Security***". In 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 327-334). IEEE, Omaha, Nebraska, USA.

[5]Anil, A., Singh, S.R. and Sarmah, R., 2018, January. "***Mining heterogeneous terrorist attack network using personalized PageRank***". In Web Intelligence (Vol. 16, No. 1, pp. 37-52).

Chapter 5

Network Embedding for HIN

The link prediction models using heterogeneous local similarity measures discussed in the previous chapter exploit explicit network structure, i.e. edge between pair of nodes. Thus, they are limited to predicting the likelihood estimate only for those nodes which are connected via a path. However, in real-word HINs, two nodes may play similar structural role even they are not connected, e.g. hubs in the social network. Therefore, this chapter explores network embedding framework that exploits latent representation of the underlying HIN apart from the exiting connectivity among the nodes. In particular, this chapter discusses some of the recent state-of-the-art network embedding models and analyses their limitations in mining a heterogeneous information network.

5.1 Introduction

Recently there is a surge in applying network embedding for addressing various tasks in network science such as classification, clustering, link prediction, community detection etc. [97, 1, 98, 103]. Network embedding aims at learning low dimensional feature vector capable of preserving network's structural characteristics [95, 97]. Further, network embedding may refer to node embedding, edge embedding, sub-graph embedding or whole graph embedding [89]. The scope of

this study (thesis) is node embedding. Majority of the network embedding models proposed previously consider homogeneous networks, i.e. network consisting of singular type of nodes and relations [98, 102, 97, 103]. However, majority of the real-world information networks and social networks are heterogeneous in nature i.e. networks consist of multiple types of nodes and relations [123]. For example, an academic bibliographic network may be represented using Author (A), Paper (P), Venue (V) (conference/journal) as nodes and different contextual relations such as Author-`writes`-Paper (AP), Author-`publishes-at`-Venue (AV), etc.

Majority of the previous studies on mining heterogeneous networks [25, 21] exploit *meta-path* [20] which is a sequence of relations between different node types. Further, symmetric meta-paths are capable of preserving heterogeneous proximity between the underlying nodes. For example, in a bibliographic network, meta-path APA gives the proximity estimate between two authors collaborating on the same paper whereas AVA represents proximity between two authors publishing at the same venue. While exploring a network, a meta-path defines a specific path the explorer should follow. Recently, meta-paths have been used to generate network embedding [1, 124] and reported to obtain promising results for various applications in network mining such as node classification, link prediction, clustering, etc. In this chapter, we systematically analyze the effectiveness of considering meta-path for generating network embedding, specifically for bibliographic network. Since, meta-path guides to explore only the partial network defined by the meta-path, it may lose some of the inherent network properties. Motivated by this, this study attempts to understand the following two important issues while considering meta-paths for generating network embeddings.

1. Does meta-path lose network information which can degrade the network embedding performance?

2. Are meta-path-based embeddings independent to the end task?

To investigate the above-discussed problems, we evaluate embeddings generated using different types of meta-paths using three state-of-the-art embedding models,

namely, (i) `Metapath2vec` [1], (ii) `Node2vec` [97], and (iii) `VERSE` [103] on Co-authorship prediction task and Author's research area classification in DBLP[1] heterogeneous bibliographic network. From various experimental observations, it is evident that meta-path-based network embedding cannot be generalized for graph-based problems of diverse nature. Further, selecting suitable node types in the underlying heterogeneous network seems to be more important than considering different meta-paths for heterogeneous network embedding.

5.1.1 Contribution

This chapter contributes in the following ways:

- Studies the effect of meta-path-based HIN embedding frameworks in solving Co-authorship prediction and Author's research area classification.
- Investigates the performance of different types of meta-path-based HIN embeddings over tasks of diverse nature, i.e. link prediction and classification.

5.2 Literature Survey

For network embedding, a majority of the initial studies attempt to map the natural graph representations like normalized adjacency or Laplacian matrix to lower dimensions by using spectral graph theory [90, 94] and various non-linear dimensionality reduction techniques [91, 92, 93]. However, these models are not scalable to large real-world networks as they exploit graph decomposition techniques at the core which require the whole matrix beforehand.

To overcome the above limitations, many network embedding models exploit a framework which first generates a neighborhood sample using a random walk or proximity measure and then leverages it to learn the node embeddings using a skip-gram [104] based neural network model [97, 98, 102]. For example, Node2vec [97]

[1]https://dblp.uni-trier.de/

uses a second order random walk to generate the neighborhood samples and learn the node embedding using skip-gram model, VERSE [103] preserves the vertex-to-vertex similarity using Personalized PageRank [105] and then exploits a single layer neural network to learn the embeddings.

All the above graph embedding models are proposed for homogeneous network. Recently, `Metapath2vec` [1] is proposed for heterogeneous network embedding which samples the node neighborhoods using a random walk guided through a meta-path. In a similar direction, study in [124] exploits the combined effect of different meta-paths of predefined length to generate node embeddings in heterogeneous network.

5.3 Network Embedding

5.3.1 Homogeneous Network Embedding

With the popularity of word2vec model using skip-gram proposed in [104] for generating word embedding from large sentence corpus, studies in [98, 97, 102] adapt skip-gram for network embedding. These network embedding frameworks exploit random walk-based sampling strategy to generate node sequences capturing node's neighborhood characteristics similar to a sentence which captures contextual relation between two words. Formally, for a given network $\boldsymbol{G(V, E)}$, network embedding using skip-gram model aims at maximizing neighborhood probability for a given node:

$$\arg\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{v} \in \boldsymbol{V}} \sum_{\boldsymbol{c} \in \boldsymbol{\mathcal{N}(v)}} \boldsymbol{log}\ \boldsymbol{p(c|v;\theta)} \tag{5.1}$$

where $\boldsymbol{\mathcal{N}(v)}$ gives the neighbors of $\boldsymbol{v}$ and $\boldsymbol{p(c|v;\theta)}$ is the conditional probability of observing neighbor node $\boldsymbol{c}$ for the given node $\boldsymbol{v}$.

5.3.2 Heterogeneous Network Embedding

For a given heterogeneous network $\boldsymbol{G(V, E, \mathcal{V}, \mathcal{E}, \phi, \psi)}$ defined in Definition (2.1), the skip-gram model defined in Equation (5.1) can be transformed into heterogeneous skip-gram model as follows [1]:

$$\arg\max_{\boldsymbol{\theta}} \sum_{v \in V} \sum_{\tau \in \mathcal{V}} \sum_{c_\tau \in \mathcal{N}_\tau(v)} log\ p(c_\tau | v; \boldsymbol{\theta}) \tag{5.2}$$

where $\boldsymbol{\mathcal{N}_\tau(v)}$ gives the neighbor nodes of $\boldsymbol{v}$ from $\boldsymbol{\tau^{th}}$ type. Furthermore, $\boldsymbol{p(c_\tau | v; \theta)}$ is defined using softmax function, i.e. $\boldsymbol{p(c_\tau | v; \theta) = \frac{exp(X_{c_\tau} \cdot X_v)}{\sum_{u \in V} exp(X_u \cdot X_v)}}$, where $\boldsymbol{X_v}$ corresponds to the embedding vector of node $\boldsymbol{v}$.

5.3.3 Meta-path-based Heterogeneous Network Embedding

The meta-path-based heterogeneous network embedding model exploits heterogeneous skip-gram defined in Equation (5.2). Further, random walks guided through meta-paths are used to generate neighborhood samples for all the nodes. In other words, random walker traverses partial heterogeneous network specific to underlying meta-path. For example, `Metapath2vec` exploits APVPA (or AVA) meta-path while generating random walk-based node sequences [1].

While `Metapath2vec` has been proposed specifically for heterogeneous network embedding, the above-discussed meta-path-based network embedding framework can be easily adapted by homogeneous network embedding methods through redefining the input network with specific meta-path. Therefore, this study further exploits two homogeneous network embedding models namely `Node2vec` [97] and `VERSE` [103] for meta-path-based heterogeneous network embedding.

5.4 Experimental Setups and Analysis

5.4.1 Experimental Dataset

This study uses DBLP bibliographic dataset (reported in [125]) covering publication information for the period between years 1968 to 2011. To generate various network embeddings using different meta-paths and to evaluate the embedding performance over different applications, we further divide the dataset into two parts; (i) from 1968 to 2008 for generating network embedding, and (ii) from 2009 to 2011 for evaluating the embeddings over different applications. This study considers three types of nodes, namely (i) Author (A), (ii) Paper (P), and (iii) Venue (V) for constructing various types of networks defined by all possible meta-paths. We construct the following four types of undirected networks from the DBLP 1968-2008 dataset.

- `AA`: It is a homogeneous unweighted co-authorship network considering only `Author` node type. Two nodes are connected if they co-author a paper.
- `APA`: It is a heterogeneous unweighted network considering `Author` and `Paper` node types. An author is connected to a paper if he/she is one of the authors of the paper.
- `AVA`: It is a heterogeneous unweighted network considering `Author` and `Venue` node types. An author is connected to a venue if he/she published a paper in that venue. This network structure is similar to the structure considered in `Metapath2vec` [1].
- `All`: It is a heterogeneous unweighted network considering all three types of nodes (`Author, Paper`, and `Venue`) and corresponding relationships between them.

Table 5.1 shows the characteristics of these experimental networks.

TABLE 5.1: Characteristics of different networks constructed over DBLP data, Training Data: 1960-2008, Testing Data: 2009-2011.

Dataset	DBLP 1968-2008								DBLP 2009-2011
	AA	APA		AVA		All			
Node Types	Author	Author	Paper	Author	Venue	Author	Paper	Venue	Author
# Nodes	162298	162298	155189	162298	621	162298	155189	621	18457
#Edges	461722	475828		326602		957856			29677

5.4.2 Experimental Setups

As mentioned above, three popular recently proposed network embedding models, namely (i) `Metapath2vec` [1], (ii) `Node2vec` [97], and (iii) `VERSE` [103] are considered to generate node embeddings. For all the models, we use the same hyper-parameter values as described in the original studies cited above. All the embedding results reported in this chapter consider 100-dimensional vector [2]. To investigate the performance of different meta-paths and their associated embedding, we evaluate the embedding quality using the following two applications.

5.4.2.1 Co-authorship Prediction:

Like the study [103], we also consider Co-authorship prediction task as a classification problem i.e., given a node pair, classify if the node pair has a co-author relation or not. To model it as a binary classification problem, we generate feature vectors representing node pairs using Hadamard operator [97, 103]. To avoid possible bias with the embedding towards the target application, we consider the DBLP 2009-2011 (non-overlapping with the embedding dataset) for generating samples for the classification task. In this sample, there are 29,677 number of co-authorship relations and 18,457 authors. We use random 80-20 split as training and test samples subjected to four different classifiers namely Gaussian Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR). To avoid over-fitting, the above setup has been repeated 10 times.

[2]While testing with different dimensions 100, 200, 300, we did not observe significant differences. We therefore consider 100-dimensional vector.

TABLE 5.2: Accuracy for Co-authorship Prediction by Classifiers for different Networks.

	Metapath2vec				Node2vec				VERSE				Combine			
Classifier	AA	APA	AVA	All	AA	APA	AVA	All	AA	APA	AVA	All	AA	APA	AVA	All
NB	0.585	0.633	0.694	0.717	0.688	0.699	0.697	0.719	0.725	0.756	0.733	0.746	0.673	0.745	0.737	0.758
RF	0.761	0.724	0.698	0.720	0.749	0.731	0.698	0.730	0.760	0.754	0.707	0.744	0.772	0.753	0.714	0.748
DT	0.683	0.654	0.628	0.644	0.678	0.658	0.632	0.657	0.688	0.674	0.642	0.678	0.699	0.673	0.645	0.678
LR	0.736	0.739	0.738	**0.766**	0.773	0.766	0.75	**0.777**	0.788	0.784	0.764	**0.796**	0.799	0.795	0.778	**0.806**

TABLE 5.3: Accuracy for Author's Research Area Classification by Classifiers for different Networks.

	Metapath2vec				Node2vec				VERSE				Combine			
Classifier	AA	APA	AVA	All	AA	APA	AVA	All	AA	APA	AVA	All	AA	APA	AVA	All
NB	0.392	0.476	0.503	0.499	0.500	0.582	0.497	0.488	0.492	0.557	0.550	0.552	0.429	0.58	0.529	0.522
RF	0.484	0.486	0.491	0.482	0.488	0.536	0.518	0.509	0.495	0.499	0.530	0.545	0.499	0.529	0.527	0.53
DT	0.442	0.439	0.439	0.428	0.436	0.481	0.472	0.449	0.445	0.440	0.476	0.490	0.456	0.471	0.474	0.495
LR	0.504	0.539	0.565	**0.566**	0.486	0.544	**0.559**	0.555	0.536	0.531	0.605	**0.624**	0.552	0.592	0.612	**0.625**

5.4.2.2 Research Area Classification:

We now investigate the quality of the embeddings for predicting author's research area. For each author in DBLP 2009-2011, we further identify (considering the `Field` attribute in [125]) the area in which author has maximum publication and consider it as the author's class label. Like Co-authorship prediction, we use similar random 80-20 split for all the classifiers and repeated 10 times.

5.4.3 Result and Discussion

Tables 5.2 and 5.3 present the Accuracy for Co-authorship prediction and Author's research area classification respectively using three network embedding models discussed above for all networks, i.e. `AA, AVA, APA`, and `All`. From Tables 5.2 and 5.3, it is observed that LR out-performs other classifiers in 93% times for Co-authorship prediction and 75% times for Author's research area classification task. Therefore, we select LR Accuracy for further analysis.

We first investigate if meta-path-based embedding loses information or not. It is evident from Tables 5.2 and 5.3 that almost all the models perform best by exploiting `All` network and show poor performance with `AA, APA`, and `AVA` networks for both tasks, i.e. Co-authorship prediction and area classification. Thus, it can be inferred that meta-path alone may be a weak representation for the network

because it does not incorporate the impacts of other relational properties while capturing node neighborhood.

Secondly, we intend to investigate if the same embedding responds coherently to different problems. From Tables 5.2 and 5.3, it is clearly visible that `APA` performs better than `AVA` for Co-authorship prediction whereas `AVA` performs better than `APA` for classifying Author's research area. This observation is true for all the embedding techniques used in this study. Thus, meta-path-based heterogeneous network embedding cannot be generalized for the tasks of different nature.

The homogeneous network `AA` and heterogeneous network `APA`, preserve similar proximity, i.e. Co-authorship between underlying pair of authors. From Table 5.2, it is evident that `AA` performs better than `APA` for Co-authorship prediction in majority of the cases. However, for Author's research area classification in Table 5.3, `APA` performs better than `AA` in almost all the scenarios. Thus, it can be inferred that meta-path-based heterogeneous network embedding may perform differently (poor or better) compared to homogeneous network embedding when subjected to tasks of diverse nature.

Among all the embedding models, `VERSE` consistently outperforms others for almost all the networks and classifiers for both Co-authorship prediction and research area classification tasks. It may be because unlike `Metapath2vec` and `Node2vec`, `VERSE` exploits a Personalized PageRank [105] capturing vertex-to-vertex similarity while generating the neighborhood sequences.

We further investigate combining all the three embeddings (`Metapath2vec`, `Node2vec`, `VERSE`) by concatenating the feature vectors. From Tables 5.2 and 5.3, it is observed that combined embedding always out-performs individual embedding for Co-authorship prediction and Author's research area classification over all the four networks.

5.5 Summary

In this chapter, we investigate the applicability of meta-paths in heterogeneous network embedding for Co-authorship prediction and Author's research area classification problems in heterogeneous DBLP database. From various experimental results, we observe that by using appropriate node and relation types, majority of the embedding methods out-perform their counter-parts exploiting meta-path-based network for both of the above-mentioned tasks. Further, it is also evident that exploiting past co-authorship relation or `APA` meta-path yields better co-author prediction in comparison to `AVA` meta-path which exploits author's publication venue. On the other hand, `AVA` meta-path contributes positively to Author's research area classification problem and have superior performance than `APA` meta-path. Thus, for heterogeneous network embedding one should carefully choose the node types, relation types, and meta-paths which can capture the network characteristics in a better way to address the underlying problem. This work has been accepted by ICADL-2018[3] and published in PReMI-2019[4].

As discussed above, selecting best meta-path is not a trivial problem. Since majority of the network embedding models require larger training time (in comparison to the topology-based models), the issue of finding optimal meta-path is more critical. Thus, there is a requirement for approaches which do not need explicit meta-paths for network embedding. In the next chapter, we propose one such network embedding-based on preserving $\boldsymbol{k}$-hop random node proximity for HIN embedding.

[3]Anil, A., Chugh, U. and Singh, S.R., 2018, August. "***On Applying Meta-path for Network Embedding in Mining Heterogeneous DBLP Network***". Accepted by International Conference on Asia-Pacific Digital Libraries (ICADL), Hamilton, New Zealand.

[4]Anil, A., Chugh, U. and Singh, S.R., 2019, August. "***On Applying Meta-path for Network Embedding in Mining Heterogeneous DBLP Network***". Published in International Conference on Pattern Recognition and Machine Intelligence (PReMI), Tezpur, India.

Chapter 6

HIN Embedding using k-hop Random Walks

The previous chapter exploits various meta-paths over state-of-the-art embedding models for heterogeneous network embedding. However, selecting the optimal meta-path is yet a critical pre-requisite. As observed in Chapter 4 that real-word HIN (i.e. heterogeneous terrorist attack network) may have large number of meta-paths and selecting the best is not an easy task. Further, from Chapter 5, it is evident that meta-path-based HIN embedding cannot be generalized and may lose network characteristics. Therefore, this chapter focuses on devising a novel network embedding model that do not require explicit meta-path information while HIN embedding.

6.1 Introduction

Network embedding using neural network over social and information networks became popular in a very recent time. Majority of the unsupervised network embedding methods use a two-step framework, *(i) generate the node sequences using a sampling method, and (ii) train a neural network on these sequences to generate embedding* [97, 1, 98].

Random walk is one of the popular and scalable sampling methods for real-world information networks. To generate network samples, early network embedding models use the first-order random walk where the transition to next node only depends on the current node [98]. However, as reported in studies [97, 126], for real-world information networks, a second-order random walk (transition to next node depends on current and previous nodes) preserves network characteristics more efficiently by capturing community dynamics and structural equivalences. Though the above random walk-based sampling methods help in generating efficient network embedding for homogeneous networks (consisting of singular type of nodes and edges), as reported in `Metapath2vec` [1] they perform poorly on heterogeneous networks (consisting of multiple types of nodes and edges). However, the main challenge of `Metapath2vec` is to select a suitable meta-path for heterogeneous networks with many types of nodes and edges.

All the embedding models discussed above exploit random walk for node sampling which is bound to consider adjacent neighbors. However, heterogeneous networks may often have neighbors of different types. For example, in a heterogeneous bibliographic network consisting of *author, paper, and venue* as nodes, two *authors* may be connected via a *paper* or a *venue*. Therefore, neighbors of different hops (distance) may potentially capture different latent characteristics. Motivated by this, this work proposes a k-hop random walk-based sampling approach (RW-k) that performs random walk on k-hop neighbors and preserves proximity of random neighbors separated by k hop or edges.

We use the RW-k based network samples for network embedding using skip-gram model [104] with negative sampling. Thereafter, we evaluate the embedding performance on Co-authorship prediction task over three heterogeneous bibliographic networks extracted from DBLP[1] and ACM[2] bibliographic databases. We compare the performance of embedding using the proposed RW-k sampling method to recently proposed embedding methods exploiting random walk-based sampling

[1]https://dblp.uni-trier.de/
[2]https://www.acm.org/

namely `Metapath2vec` [1], `Node2vec` [97] and `VERSE` [103]. From various experiments, it is evident that RW-k outperforms the state-of-art network embedding methods and improves the performance of Co-authorship prediction significantly.

6.1.1 Contribution

The major contributions of this chapter are:

- A novel random walk-based sampling approach (RW-k) suitable for mining heterogeneous information networks.
- Empirical evaluation of RW-k based heterogeneous network embedding on Co-authorship prediction task over three different heterogeneous bibliographic networks.

6.2 Node Embedding

Similar to the Section 5.3.1, this study also model the node embedding problem by solving an optimization problem. In particular, we maximize the log probability of appearance of a sequence of nodes $\mathcal{N}_v$ preserving k-hop proximity characteristics for the given node v using Equation (5.1).

To generate the node sequences $\mathcal{N}$, several sampling approaches have been used in past. One of the popular sampling approach is based on the random walk process that captures the immediate neighborhood. However, in context to the HIN, traversing immediate neighbors results to sample different types of nodes closely which may degrade the quality of node embedding. Therefore, in this study, we propose a random walk-based sampling approach RW-k which generates the node sequences for a given source node in which two adjacent nodes are separated by some hops k. Further, RW-k is capable of capturing k hop proximity helpful in generating more meaningful node sequences leading to better node embedding.

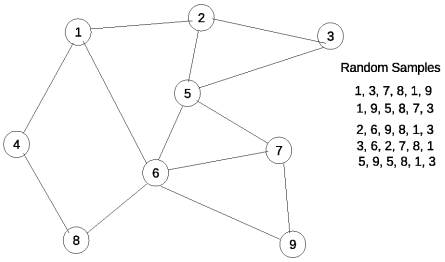


FIGURE 6.1: Network Sampling Using RW-2.

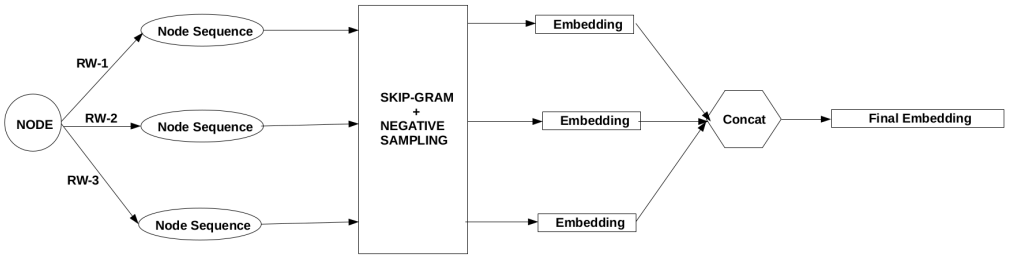


FIGURE 6.2: Node embedding using concatenation of different values for k in RW-k.

6.3 Network Sampling using k-hop Random Walk (RW-k)

In network science, a hop can be understood as a quantity such as distance, time, etc. required to make a visit from the given source node to destination node. For example, in Figure 6.1 node pair $(4, 1)$ is at the 1-hop while 2-hop distance is required for visiting node **2** from **4**. As a special case, each node is separated from itself by 2-hop distance.

For a given Graph $G(V, E)$, let $\mathbf{A}$ be the adjacency matrix and π be the corresponding transition matrix, where the transition probability from the node u to v is defined as

$$\pi_{uv} = \begin{cases} \frac{\mathbf{A}_{uv}}{\sum_{w \in V} \mathbf{A}_{uw}} & if(uv) \in E \\ 0 & Otherwise \end{cases} \tag{6.1}$$

where $\sum_{v \in V} \pi_{uv} = 1$.

Now for a given hop size $k \in \mathbb{Z}$, RW-k generates a node sequence starting from source u such that every random node in the sequence are separated by k hop distance. Suppose $\pi^k_{(u.)}$ gives the transition probability from source u to other destination nodes separated by k hops, then π^k_{uv} can be defined as

$$\pi^k_{uv} = \sum_{w \in V} \pi^{k-1}_{uw} \pi_{wv} \tag{6.2}$$

where $\sum_{v \in V} \pi^k_{uv} = 1$. This can further be realized using transition matrix π that k-hop transition probability between all pairs of nodes i.e. $\pi^k = \pi^{k-1}\pi$. It can be easily seen as π^k is a stochastic matrix. Figure 6.1 shows example node sequences for hop size 2 using RW-2 which samples nodes having two-hop distance in between them.

We exploit the skip-gram model with negative sampling discussed in Section 6.2 for node embedding using sequences generated by RW-k with different values of k, i.e. $k = 1$, $k = 2$, and $k = 3$. Further, selecting an optimal value for k is dependent on the underlying network's characteristics, such as network schema, distribution of node types, network density, etc.

6.3.1 Aggregate k-hop Embedding

The node embedding generated using different values of k in RW-k do not capture the network characteristics described by intermediate k. Therefore, to capture the characteristics of each intermediate RW-k, we propose concatenation of the node embedding obtained by different RW-k for each node. Figure 6.2 presents the concatenation process.

Table 6.1: Network Characteristics.

Sl. No.	Dataset	Node	Edge	Timeline
1	DBLP-All	318108	957856	1968-2008
2	DBLP-conf	267846	751479	1934-2016
3	ACM-conf	120863	270363	1984-2015

6.4 Experimental Analysis

6.4.1 Dataset

We use three heterogeneous bibliographic networks for evaluating the quality of network embedding on predicting future Co-authorship relations. We consider Author, Paper, and Venue as node types and corresponding relationships among these node types to generate heterogeneous bibliographic networks.

1. **DBLP-All:** DBLP bibliographic information from 1968 to 2011 reported in the study [125]. It contains all types of papers i.e. conference, journal, book, etc.

2. **DBLP-conf:** DBLP bibliographic information for the different categories of conferences [3] from 1934-2017 downloaded from Aminer dataset [4].

3. **ACM-conf** ACM bibliographic information for the different categories of conferences used in DBLP-conf from 1984 to 2016 downloaded from Aminer dataset.

We consider DBLP-All (1968 to 2008), DBLP-conf (1934 to 2016), and ACM-conf (1984 to 2016) for training the network embedding model. Table 6.1 presents the network characteristics.

[3]http://www.conferencelist.info/target.html
[4]https://aminer.org/citation

TABLE 6.2: AUC Score for Co-authorship Prediction by Classifiers for different Networks. Metapath2vec uses Author-Venue-Author meta-path as suggested by [1].

Dataset	Classifier	Metapath2Vec	Node2Vec	VERSE	RW-1	RW-2	RW-3	Concat(RW-1, RW-2, RW-3)
DBLP-All	DT	0.628	0.657	0.678	0.647	0.673	0.673	**0.679**
	NB	0.694	0.719	**0.746**	0.69	0.718	0.713	0.718
	RF	0.698	0.73	0.744	0.721	0.745	0.75	**0.752**
	LR	0.738	0.777	0.796	0.766	0.78	0.792	**0.804**
DBLP-conf	DT	0.594	0.625	0.62	0.624	0.635	0.647	**0.651**
	NB	0.627	0.685	**0.703**	0.666	0.689	0.684	0.683
	RF	0.664	0.695	0.692	0.683	0.706	0.706	**0.71**
	LR	0.691	0.734	0.738	0.73	0.751	0.753	**0.762**
ACM-conf	DT	0.615	0.622	0.6	0.607	0.611	0.629	**0.635**
	NB	0.621	0.639	0.653	0.628	0.637	0.637	**0.664**
	RF	0.671	0.686	0.664	0.681	0.679	0.697	**0.699**
	LR	0.656	0.701	0.697	0.696	0.703	**0.716**	0.714

6.4.2 Experimental Result and Discussion

To generate the network samples using RW-k, we investigate with the different values of k i.e. 1, 2, 3,..., etc. We observed that $k > 3$ does not improve the embedding quality significantly. Thus, we consider k only upto 3 and generate the samples (maximum of **100** in length) by iterating 30 times for any given node in the network[5]. Now, for network embedding, we exploit the skip-gram model with negative sampling on these samples.

For evaluating the quality of embedding we use future Co-authorship relations appeared in DBLP-All (2009 to 2011), DBLP-conf (2017), and ACM (2016) as test links. Similar to the studies [97, 103] , we map the Co-authorship prediction task from link prediction to a binary classification task. Thereafter, we generate an equal number of negative test links on the nodes appeared in the above test links. Similar to studies [97, 103], we generate the edge feature using Hadamard operator on the node embeddings. We use four state-of-art classifiers namely, Decision Tree (DT), Naive Bayes (NB), Random Forest (RF), and Logistic Regression (LR) on a random 80:20 train-test split over combined test links (actual test links with randomly generated negative test links). Thereafter, we repeat the same setup for 10 times and the average AUC is considered to assess the performance of co-authorship prediction. We compare the performance of Co-authorship prediction

[5]We investigated with higher iteration value i.e. 40, 50, etc. and observed that higher than 30 iterations do not show much difference.

by network embedding exploiting RW-k with recently proposed network embedding models namely `Node2vec` [97], `Metapath2vec` [1], `VERSE` [103] exploiting first order random walk, second order random walk, and personalized PageRank [118] respectively for sampling the network. We consider the default values for all the hyper-parameters for the above discussed baselines as mentioned in their original papers and generate 100 dimensional embedding vector for each node.

From Table 6.2, it is evident that network embedding exploiting RW-k outperforms the state-of-art baselines in 83% times. Therefore, it can be inferred that in heterogeneous information networks, capturing neighborhood characteristics separated by some hops helps in generating better network embedding. We observe that Logistic Regression (LR) outperforms other classifiers for all the three networks. Therefore, we present the remaining analysis based on the performance of LR for future Co-authorship prediction task. An improvement of 9%, 3%, and 1% is achieved for Co-authorship prediction in DBLP-All using RW-k over Metapath2vec, Node2vec, and VERSE respectively. Furthermore, for DBLP-conf and ACM-conf, RW-k improves the performance approximately by 9%, 2%, and 3% in comparison to Metapath2vec, Node2vec, and VERSE respectively.

From Table 6.2, it is observed that concatenation of different RW-k (i.e. RW-**1**, RW-**2**, and RW-**3**) improves the performance of Co-authorship prediction when compared to individual RW-k. Further, it is also visible that RW-2 gives a better result than RW-1 and RW-3 gives a better result than RW-2 for all the networks. However, the improvement using RW-3 over RW-2 is not as significant when compared to RW-2 over RW-1. This shows that capturing neighborhood characteristics with small hops can yield better network embedding though increasing the hop size does not show significant improvement. Hence, it can be inferred that different hop sizes for RW-k preserves rich network characteristics and should be selected carefully.

Among all the above network embedding models `Metapath2vec` performs worst. The probable reason for this may be the loss of network information while sampling using meta-path.

6.5 Summary

This chapter studies the problem of network sampling for heterogeneous network embedding. We propose a k-hop random walk-based sampling approach (RW-k) that generates more meaningful node sequences. These node sequences are further used for network embedding using skip-gram with negative sampling. The efficacy of the proposed sampling approach in network embedding is evaluated on Co-authorship prediction task over three heterogeneous bibliographic networks and compared to suitable baselines. We observe that the proposed sampling approach out-performs the baselines in a majority of the cases.

The limitation of this work is that it requires a hyper-parameter k as hop size. To alleviate this limitation, learning the optimal hop size from supervised knowledge of the underlying network may be a future direction. This chapter has been accepted by ICADL-2018[6] and published by CoDS-COMAD-2019[7].

All of the HINs considered in the previous chapters (or any real-world HIN) consist of multiple types of objects. Further, these multiple types of objects may have different number of instances. Thus, class imbalance seems to be an inherent characteristics of HIN. In the next chapter, we study the effects of class imbalance over network embedding in mining a HIN.

[6]Anil, A., Ladhar, A., Singh, S. and Chugh, U., Singh, S.R., 2018, August. "***Network Sampling Using k-hop Random Walks for Heterogeneous Network Embedding***". Accepted by International Conference on Asia-Pacific Digital Libraries (ICADL), Hamilton, New Zealand.

[7]Anil, A., Singhal, S., Jain, P., Singh, S.R., Ladhar, A., Singh, S. and Chugh, U., 2019, January. "***Network Sampling Using k-hop Random Walks for Heterogeneous Network Embedding***". In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (pp. 354-357). ACM, Kolkata, India

Chapter 7

Effect of Class Imbalance in HIN Embedding

A majority of the real-world HINs consist of multiple types of objects (or nodes) having irregular number of instances. For example, bibliographic HINs discussed in the previous two chapters consist of large number of nodes from Author or Paper types when compared to Venue node types; terrorist attack network in Chapter 2 consists of large number of nodes from city type when compared to region node type. Thus, in real-world HINs a skewed distribution of instances under different node types is ubiquitous. The skewed distribution of different object types is often seen as class imbalance problem in machine learning and data mining community. Although class imbalance have been extensively studied in past, yet have not been explored much for HINs. As class imbalance seems to be an intrinsic characteristics of HIN, this chapter attempts to understand the effect of class imbalance on HIN embedding subjected to two real-world tasks in regards to bibliographic data namely, Co-authorship prediction and Author's research area classification.

7.1 Introduction

Majority of the real-world data are heterogeneous in nature, consisting of multiple types of objects and relations. For example, bibliographic data often consists of different objects like Author (A), Paper (P), Venue (V) and connected with different relationships such as Author **writes** Paper (A-P), Paper **published-in** Venue (P-V), etc. A natural extension to the heterogeneous data is Heterogeneous Information Network (HIN) where different types of objects are represented by different classes of nodes and different types of relations are represented by different classes of edges [21]. Recently, neural network-based network embedding methods have emerged as a popular choice of generating network features which can be used for various network mining tasks (on homogeneous networks [98, 102, 101, 95, 97, 96] and heterogeneous information networks [127, 1, 124, 128]). Unlike homogeneous networks, a HIN is naturally subjected to class imbalance i.e. different number of instances in different classes of nodes. In recent past, the problem of class imbalance has been briefly mentioned in regards to heterogeneous network embedding in the studies [1, 124, 128]. However, a detailed investigation on the effect of class imbalance in mining HIN has not been explored. While HIN offers rich and complex semantics, class imbalance seems to be the inherent characteristic of HIN. Motivated by the above characteristics, this study focuses on investigating the effect of class imbalance on various network mining tasks over a bibliographic HIN.

Class imbalance is ubiquitous in case of real-world HIN. However, it is practically difficult to define a unique measure for class imbalance. As discussed by Chawla et al. [129], in real-world datasets the amount of class imbalance between large to small classes may be equal to 100 to 1 or more. Since a HIN may have more than two classes of nodes, we define the amount of class imbalance by averaging the different ratios of node classes corresponding to different relations in HIN (discussed in Section 7.3). Consequently, different HINs (differ in terms of node types and edge types) constructed for the same data may have a different amount of class imbalance and can affect the performance of underlying models designed for solving various graph-based problems. As this chapter focuses on studying

the effect of class imbalance on network embedding over HIN, we first attempt to understand the following question: *how class imbalance affect HIN embedding and its applications on various network mining tasks?*

Majority of the network embedding methods for HIN exploit a set of supervised paths or meta-paths capturing particular sub-structure of the underlying HIN to overcome class imbalance effects [1, 124]. Although meta-path-based network embedding results in better embedding quality, selecting best meta-path is not a trivial task for large HIN [128]. For example, a knowledge graph is a HIN which may consist of a large number of meta-paths and selecting the best meta-path requires a careful investigation. Study in [128] uses a probabilistic approach which decides the transition probability by estimating jump or stay probability. Though, no meta-path is required with this model, it inherently forces the model to follow a particular node sequences similar to meta-path-based strategies. Considering the extent of using meta-path-based methods, we further investigate the above question from the perspective of meta-paths on HIN embedding.

As discussed above meta-path-based network embedding models capture network characteristics for a supervised sub-structure of the underlying HIN. However, considering particular sub-structure may affect the performance for the underlying problems. For example, study in [130] shows that HIN considering Author, Paper, and Venue results in better Co-authorship prediction whereas for Author's research area classification HIN with Author and Venue gives better accuracy. Therefore, in a similar direction of [130], we further study the importance of node type selection on network embedding for solving network mining problems of different nature i.e. link prediction and classification.

To address the above-discussed research questions related to the class imbalance in HIN, we exploit two state-of-the-art network embedding models namely, DeepWalk [98] and VERSE [103] over DBLP [1] bibliographic information network. Further, we exploit the node embeddings generated from these models as features in

[1]https://dblp.org

solving two problems namely *(i) Co-authorship prediction*, and *(ii) Author's research area classification*. We generate several variants of this bibliographic HIN and with suitable experiments present the observations on the effects of class imbalance over heterogeneous network embedding. We observe that class imbalance in HIN affect the mining performance differently for the different network mining problems and may lead to subjective speculations.

7.1.1 Contribution

This chapter has the following major contributions:

1. We present a detailed study on the effect of class imbalance in mining heterogeneous bibliographic network for different network mining problems.
2. We present a comparative study which analyses the effect of class imbalance in heterogeneous information network embedding with varying amount of class imbalance on two applications namely, (i) Co-authorship prediction and (ii) Author's research area classification.
3. Motivated by the applications of random walks in unsupervised network embedding, we further present a comparative study for the effect of class imbalance on inter-class and intra-class node ranking problem.

7.2 Literature Survey

Although none of the earlier works on heterogeneous network embedding study the effect of class imbalance, they agree on a common perception, i.e. *class imbalance may bias the node embeddings towards highly visible node types* [1, 124, 131, 128]. To address class imbalance while HIN embedding, a majority of the previous studies exploit a meta-path-based strategy which guides the underlying model through supervised sequences of node types [1, 124, 131]. In particular, meta-path-based

network embedding models first select the best meta-path from the known heuristics and redefine the network neighborhood (node sequences) by iterating a random walk over the chosen meta-path. Thereafter, the network neighborhood is passed to SkipGram like model to generate node embeddings. For example, Metapath2vec [1] exploits A-P-V-P-A (representing two authors publishing papers at the same venue) as meta-path suggested from previous studies [123, 132, 133] for embedding a bibliographic HIN. It is clearly visible that the meta-path-based strategy requires domain-specific prior knowledge and may become a bottleneck for HINs having large number of node types. Further, study in [134, 130] reports that meta-path selection may be task-specific and the node embeddings cannot be generalized.

Task-specific meta-path-based network embedding models often combine a set of predefined meta-paths (dedicated to the end task) to achieve a better quality of embeddings [135, 134, 136]. For example, Shang et al. in the study [135] proposed `ESim` which exploits the combined effects of different user-defined meta-paths for similarity search. In a similar direction, Shi et al. proposed `HERec` [134], which fuses the multiple embeddings w.r.t. different meta-paths for recommendation task. Further, in the study [136], Chen et al. combined various task-guided meta-paths and proposed a path-augmented heterogeneous graph embedding for author identification task. However, selecting a set of task-specific meta-paths still cannot be solved trivially and may not be scalable for large graphs. Another approach for selecting the set of meta-paths followed by recent works is by setting a threshold on the length of meta-paths. Exploiting the meta-path's length threshold parameter, `HIN2vec` [124] leverages the combined effects of different meta-paths (smaller than the threshold in length) for guiding the random walk while learning node embedding. Similarly, `HINE` [131] guides random walk exploiting various meta-paths shorter in length to learn a node embedding preserving meta-path-based proximity. However, from [131], it is observed that node embedding is sensitive to the length of the meta-paths considered and the extended computation for combining multiple meta-paths further increase the training time.

As an alternative to meta-path-based approaches for addressing the class imbalance in HIN embedding, Hussein et al. proposed `JUST` [128] which guides a random walker according to a transition probability estimated for jumping from one node type to another node type or staying on the similar node type. Although `JUST` do not require meta-paths explicitly, it guides the random walker approximately in a similar fashion by tuning different parameters for jump and stay which can be different for different datasets and needed extended empirical investigation.

Although all the above studies for HIN embedding exploit sub-structures to reduce the effects of class imbalance, class imbalance may still be inherently present in the selected sub-structures. Therefore, this chapter studies the effect of different degree of class imbalance present in HIN for solving diverse nature of graph-based problems (i.e. link prediction and classification).

7.3 Class Imbalance

Real-world datasets often consist of different number of instances in different types of data points. Suppose the underlying dataset has two types of data points categorized as either majority class $\mathcal{C}_{maj}$ or minority class $\mathcal{C}_{min}$, then amount of class imbalance is equal to the ratio between $|\mathcal{C}_{maj}|$ to $|\mathcal{C}_{min}|$ classes. Previous researches in data mining and machine learning report class imbalance as a major bottleneck in solving problems such as text classification [137] and speech recognition [138]. It is because the underlying learning model may get biased towards highly visible data-points while training and may predict incorrect results. Realizing the possible effects of class imbalance, multiple strategies were proposed to reduce the class imbalance by under-sampling the majority class [139] and over-sampling the minority class [140]. Although the reduction of class imbalance by these approaches help the predictive models to yield stable and better performance, these are limited to feature-based machine learning strategies or non-networked datasets. Class imbalance in HIN is more complex when compared to non-networked data because under-sampling or over-sampling may distort the

structural properties of the underlying network and may not be suitable for addressing the issue.

7.3.1 Class Imbalance in Heterogeneous Information Network

As discussed in Section 7.1, we define the class imbalance in HIN by aggregating the ratio of number of nodes in majority class to minority class given that there exists an edge between these two node classes.

Definition 7.1 (Class Imbalance in HIN)**.** The amount of class imbalance can be defined as

$$CI = \frac{1}{|\mathcal{E}|} \sum_{<u,v>\in\mathcal{E}} \left\lfloor \frac{|u|}{|v|} \right\rfloor,$$

where $\mathcal{E}$ is the edge type, u and v are node types in $\mathcal{V}$ forming a particular edge type. The cardinality of u i.e. $|u|$ denotes the number of nodes of the type u. The node type with larger cardinality is considered in the numerator.

7.4 Dataset and Experimental Analysis

This section begins with describing dataset considered in this study which is further used to construct different heterogeneous information networks. Thereafter, it presents the experimental setups and experimental observations in regards to the effect of class imbalance while solving different network mining problems exploiting network embeddings as features.

7.4.1 Dataset

This study uses DBLP bibliographic database consisting of records for papers published from 1968 to 2011 [125]. In particular, it consists of five main objects

defining bibliographic data such as authors, paper, venue, topic, and field. This dataset has 162,298 authors, 621 venues, 26 topics, and 7 fields related to 155,189 papers.

7.4.2 Constructing Heterogeneous Information Network

Generalizing a standard HIN schema for datasets from different domains is not a trivial task and previous studies have exploited different types of network schema such as star [141, 142, 72, 143, 144, 73], bipartite [145, 146, 147, 148], dynamic [149, 114], clique [150, 151], etc. In this study, we construct two undirected and unweighted HINs for the above discussed DBLP data. All the objects namely, Author (A), Paper (P), Venue (V), Topic (T), and Field (F) have been considered as node types whereas these two HINs differ in considering edge types. The first one follows the HIN schema similar to [150, 151] based on clique-based architecture and considers all possible (ten types) edges. Therefore, we refer this HIN as DBLP-Clique (DBLP-C) in the subsequent parts of this study. The second HIN follows a user-specific schema which considers only four types of edges namely, A-P, P-V, P-T, and V-F. Therefore, we refer this HIN as DBLP-Personalized (DBLP-P) in subsequent parts of the study. Furthermore, we use the HIN from the interval 1968-2008 for generating node embedding and 2009-2011 network is used for testing the embedding performance.

7.4.3 Experimental Setup

Since this chapter studies class imbalance in HIN over network embedding, we further generate different variants (in terms of class imbalance) of above discussed HINs i.e. DBLP-C and DBLP-P. Let $\boldsymbol{G(V, E, \mathcal{V}, \mathcal{E}, \phi, \psi)}$ defined in Definition (2.1) represents DBLP-C or DBLP-P and similarly, $\boldsymbol{G'(V', E', \mathcal{V}', \mathcal{E}', \phi, \psi)}$ defines the underlying HIN variant, then $\boldsymbol{V' \subseteq V}$, $\boldsymbol{E' \subseteq E}$, $\boldsymbol{\mathcal{V}' \subseteq \mathcal{V}}$, and $\boldsymbol{\mathcal{E}' \subseteq \mathcal{E}}$ where, $\boldsymbol{|\mathcal{V}'| \geq 2}$. Table 7.1 presents the characteristics of HIN variants considered in this study for DBLP-C and DBLP-P. We refer these HIN variants

TABLE 7.1: Different Variants of DBLP-Clique (DBLP-C) and DBLP-Personalized (DBLP-P) constructed on the basis of edge types. APVTF in DBLP-C and DBLP-P represents entire DBLP-C and DBLP-P respectively. Class imbalance is estimated using definition 7.1.

HIN	HIN-Variants	Class Imbalance	#Nodes	#Edges
DBLP-C	**AF**	23185	162305	210980
	APF	15118	317494	842080
	ATF	9810	162331	461346
	APTF	9595	317520	1247635
	AVF	7845	162926	537527
	APVF	7659	318115	1324053
	AT	6242	162324	250185
	APVTF	5819	318141	1736597
	AVTF	4967	162952	794882
	APT	4070	317513	881202
	AVT	2175	162945	583100
	APVT	2124	318134	1369543
	AV	261	162919	325926
	APV	170	318108	957180
	AP	1	317487	475828
DBLP-P	**APT**	2985	317513	631017
	APVT	2073	318134	786443
	APVTF	1577	318141	787064
	APV	125	318108	631254
	APVF	113	318115	631875
	AP	1	317487	475828

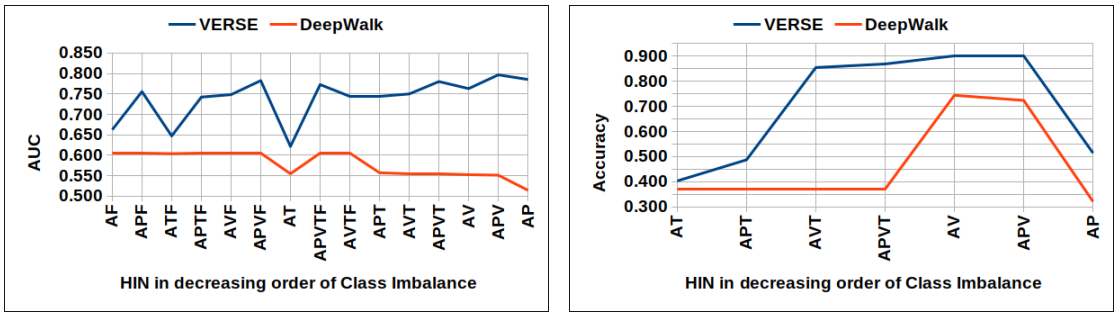
by their node types. For example, APF in Table 7.1 refers to a HIN considering Author, Paper, and Field as node types. It should be noted that AP is similar for DBLP-C and DBLP-P as there could be only one undirected edge type between Author and Paper. The HINs in Table 7.1 are used for generating 128-dimensional node embedding by DeepWalk and VERSE. We consider all the hyper-parameters same as discussed in the papers for corresponding models [98, 103]. Thereafter, we evaluate the embedding performance on the future Co-authorship prediction and Author's research area classification tasks.

7.4.3.1 Co-authorship Prediction

As discussed in Section 7.4.2, we use 29,677 number of future co-authorship instances appeared in the interval 2009 to 2011 as test edges. The network embedding models considered in this study (DeepWalk and VERSE) output node features not the edge features. Therefore, similar to studies [97, 103], we generate edge features using Hadamard product of the corresponding node features. Further, we randomly generate an equal number of negative test edges and similar to [97, 103] we solve the Co-authorship prediction problem as a binary classification problem. As noted from previous studies on this dataset [130, 152], logistic regression classifier performs better than other classifiers, we use logistic regression as a classification model. We do five-fold cross validation and report the performance of Co-authorship prediction using Area Under ROC score.

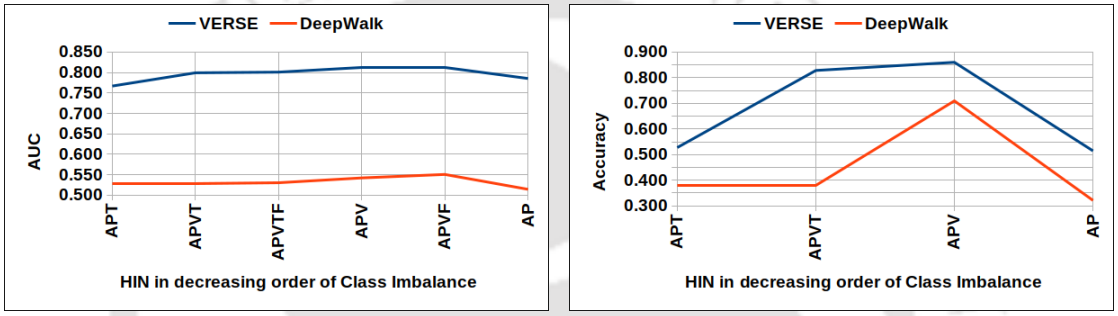
7.4.3.2 Author's Research Area Classification

The node embedding generated by the network embedding models can be used as node features. Thus, these node features can be used for the problems like node classification, node clustering, etc. This study further exploits the node embeddings for Author's research area classification task. We use the field attribute in the underlying dataset as author's research area. Further, all the HINs not considering field (F) in Table 7.1 are considered for this task. In this data, we have authors working in multiple fields. However, we label an author's area in which field he/she has published the maximum number of papers. In particular, we have 7 different fields or classes. Thus, we solve a multiclass classification problem to classify authors for their research area. We use logistic regression (one-vs-rest) as a classification model with five-fold cross validation. We report the classification performance using Accuracy score.



(a) Co-authorship Prediction (b) Author's Research Area Classification

FIGURE 7.1: Performance of network embedding over different HINs in DBLP-C for Co-authorship Prediction and Author's Research Area Classification



(a) Co-authorship Prediction (b) Author's Research Area Classification

FIGURE 7.2: Performance of network embedding over different HINs in DBLP-P for Co-authorship Prediction and Author's Research Area Classification.

7.5 Experimental Observation

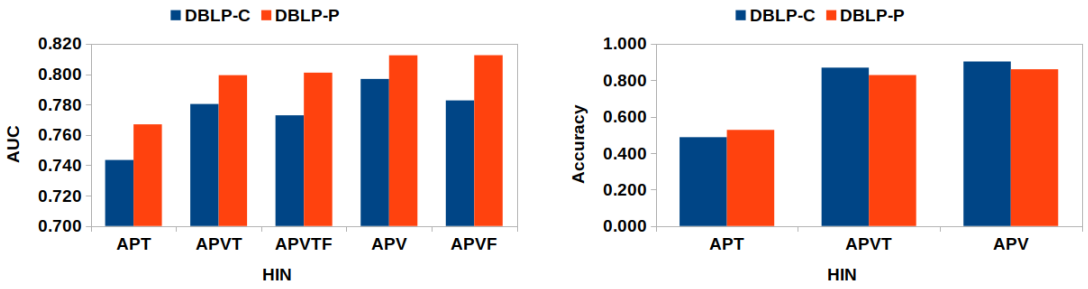
As discussed above, this work studies the effect of class imbalance on heterogeneous network embedding by solving two tasks namely, Co-authorship Prediction and Author's research area classification. We first present the effect of decreasing class imbalance on embedding quality. Thereafter, we compare the embedding efficiency between DBLP-C and DBLP-P to understand the effect of network schema and class imbalance in combination. As discussed above, majority of the heterogeneous network embedding models are based on meta-paths, we further study the embedding performance by the embedding models guided through supervised meta-paths and compare the performance with their non-meta-path based counterparts. Finally, we present an observation on the importance of considering node types while mining a HIN using network embedding.

7.5.1 Network Embedding and Decreasing Class Imbalance

To investigate the correlation between the amount of class imbalance of a HIN and its embedding for solving a given task, in Figures 7.1 and 7.2 we present the embedding performance on Co-authorship prediction and Author's research area classification over different HINs with different degree of class imbalance. It is evident from the figures that decreasing the class imbalance has positive as well as negative effects in terms of performance for both of the above-mentioned tasks. For Co-authorship prediction task in DBLP-C (Figure 7.1(a)), APV performs best while AT performs worst. However, HINs APV and AT in DBLP-C have lower class imbalance than 86% and 40% of the HIN variants respectively. For Author's classification task in DBLP-C (Figure 7.1(b)), APV and AV perform better than other HINs whereas AT performs the worst. Further, out of all HIN variants considered for this task in DBLP-C, APV and AV are ranked lower than 71%, and 57% respectively in terms of class imbalance whereas AT has the highest amount of class imbalance.

Now, for Co-authorship prediction in DBLP-P (Figure 7.2(a)), we observe APVF and APV perform better than other HINs whereas APT performs worst. Further, it is clearly visible from Figure 7.2(a) that APVF and APV are ranked lower than 67% and 50% HINs respectively in terms of class imbalance while APT has the maximum amount of class imbalance. For Author's classification task in DBLP-P (Figure 7.2(b)), APV performs the best whereas AP performs the worst. Further, in terms of class imbalance APV is ranked lower than 50% of the HINs whereas AP has the lowest amount of class imbalance. From the above experimental observations, it can be inferred that the degree of class imbalance in HIN and embedding performance on underlying tasks may not have positive correlation.

As observed in Figures 7.1 and 7.2 VERSE performs better than DeepWalk in all the cases, we consider only the VERSE embedding for rest of the experimental analysis.



(a) Co-authorship Prediction

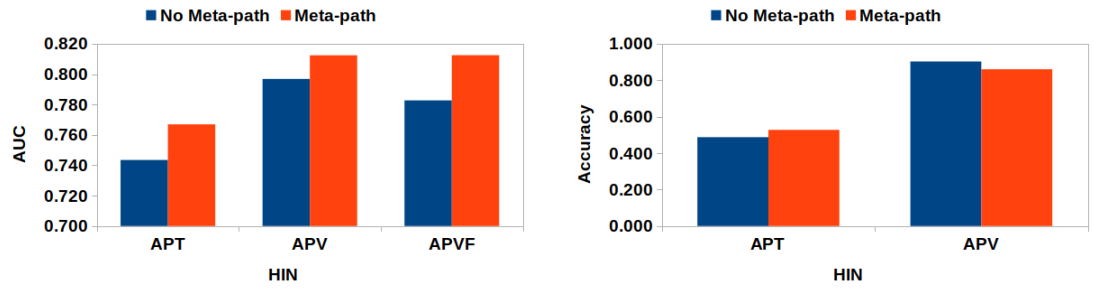
(b) Author's Research Area Classification

FIGURE 7.3: Performance of network embedding over HIN variants of DBLP-C and DBLP-P having similar type of nodes on Co-authorship Prediction and Author's Research Area Classification tasks.

7.5.2 Network Schema, Class Imbalance, and Network Embedding

As discussed in Section 7.4.1, several types of schema have been adopted by previous studies in mining HIN. Since this study defines the class imbalance corresponding to the relations present in HIN, different HIN schema will have a different amount of class imbalance. Therefore, we now attempt to study the effect of class imbalance on network embedding exploiting HINs constructed by considering two schemas namely, (i) clique-based i.e. DBLP-C, (ii) user-personalized i.e. DBLP-P.

It is evident from Table 7.1 that all the HIN variants of DBLP-P have lesser class imbalance than their corresponding counterparts in DBLP-C (excluding AP as it is identical for DBLP-C and DBLP-P). From Figure 7.3, it is evident that all the HINs in DBLP-P perform better than their counterparts in DBLP-C for Co-authorship prediction task. However, majority of the HINs in DBLP-P perform poor than their counterparts in DBLP-C for Author's research area classification task. Thus, it can be inferred that class imbalance may have different effects on network embeddings over different HIN schemas for solving different tasks.



(a) Co-authorship Prediction (b) Author's Research Area Classification

FIGURE 7.4: Performance comparison for network embedding based on meta-path vs non-meta-path counterparts for Co-authorship prediction and Author's research area classification in HINs having similar type of nodes. Network embedding based on meta-path are guided by APTPA, APVPA, and APVFVPA meta-paths for APT, APV, and APVF HINs respectively.

7.5.3 Class Imbalance, Meta-paths, and Network Embedding

Network embeddings over HINs in DBLP-P, namely APT, APVF, APV, and AP follow supervised meta-paths. For example, any network embedding models are guided to follow APTPA meta-path for APT, APVFVPA meta-path for APVF, and APVPA meta-path for APV HINs. Further, it is intuitive that network embedding using meta-path over these HINs would have lower class imbalance than their counterparts which do not follow any meta-path. Therefore, to study the effects of class imbalance on network embedding guided through a supervised meta-paths we compare the embedding quality of APT, APVF, APV in DBLP-P against their corresponding counterparts (that do not consider meta-path) in DBLP-C. Since these meta-path-based HINs belong to DBLP-P, the experimental observation is consistent with the Section 7.5.2 i.e. meta-path-based methods give better performance for Co-authorship prediction however they show mixed performance for Author's research area classification as shown in Figure 7.4. Thus, it can be inferred that meta-path reduces the class imbalance but can perform differently in different scenarios.

7.5.4 Importance of Node Types

A HIN consists of several types of nodes and edges. However, there may be some of the node types and edge types more important and may play a major role in mining a HIN efficiently. Therefore, this chapter now studies the importance of node types in mining HIN using network embedding. From Figure 7.1(a), it is clear that Co-authorship prediction is more accurate when the combination of paper and venue are considered in HIN for DBLP-C. For example, HINs by considering paper and venue along with any choice of other node types give 90% times better result than the HINs which do not consider paper and venue. However, for Author's classification task in Figure 7.1(b) we observe that paper does not contribute much and considering venue alone with other node types gives a better result than HINs which do not consider the venue. Moreover, it is also seen that combination of paper and venue gives comparable results to considering venue for Author classification task. In the case of DBLP-P schema, any variant of HIN must have to consider paper node type. Thus, we assess the importance of considering the venue along with other possible node types for both of the above-mentioned tasks for DBLP-P. From Figure 7.2, it is evident that considering the venue in HINs improve the performance for both tasks i.e. Co-authorship prediction and Author's classification when compared to HINs which do not consider venue class. Thus, it can be inferred that selecting appropriate node types is more critical than focusing on the degree of class imbalance in mining HINs for different applications.

7.6 Effect of Class Imbalance on Centrality Estimation

Majority of the unsupervised network embedding methods exploit random walks to generate network neighborhood [98, 102, 101, 95, 97, 96, 127, 1, 124, 128]. Further, random walks have also been used in defining various centrality measures such as PageRank and Personalized PageRank [105], etc. Therefore, this section attempts

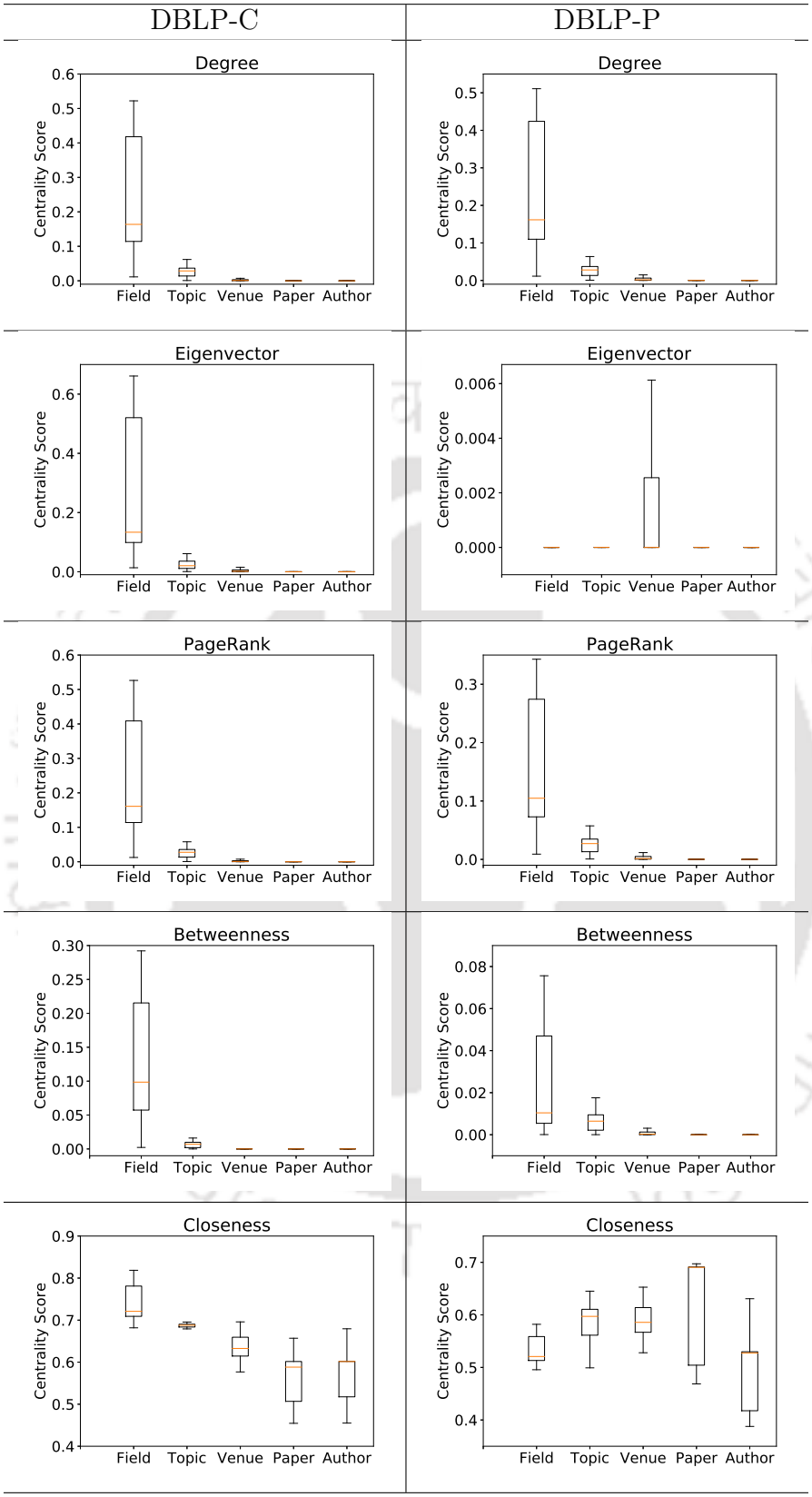


FIGURE 7.5: Inter-class centrality distribution over DBLP-C and DBLP-P exploiting Degree, Eigenvector, PageRank, Betweenness, and Closeness as centrality measures for different node classes arranged in increasing order of their cardinality.

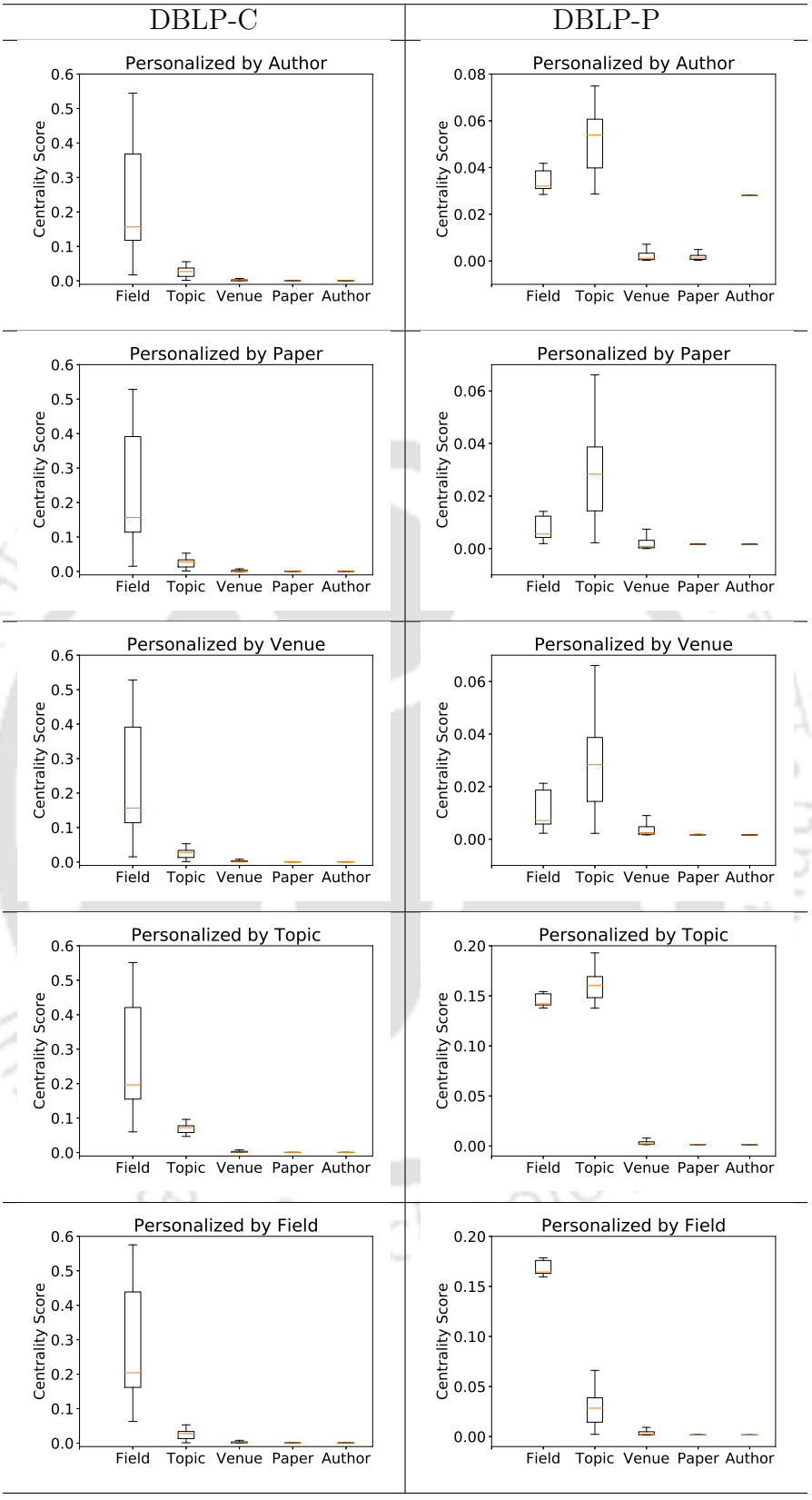


FIGURE 7.6: Inter-class centrality distribution over DBLP-C and DBLP-P by Personalized PageRank parameterized by different node types for different node classes arranged in increasing order of their cardinality.

to understand the effects of class imbalance on various centrality measures while ranking nodes in the above discussed HINs (i.e. DBLP-C and DBLP-P). In particular, we use the following state-of-the-art centrality measures: *(i) Degree, (ii) Betweenness, (iii) Closeness, (iv) Eigenvector, (v) PageRank, and (vi) Personalized PageRank (PPR).* At first, we study the effect of class imbalance in HIN on inter-class centrality distribution. Thereafter, we examine the intra-class centrality correlations for Author, Paper, and Venue nodes subjected to APVTF and APV HINs for both DBLP-C and DBLP-P.

7.6.1 Inter-class Centrality Distribution

From Figure 7.5, it is evident that for DBLP-C, all the node classes having lower number of instances or cardinality achieve larger centrality scores whereas nodes belonging to classes with higher cardinality are ranked lower. Further, in the case of DBLP-P we observe similar trends for Degree, PageRank and Betweenness whereas a contrastive result is observed for closeness and Eigenvector centrality measures. Thus, it can be inferred that HINs based on clique schema (DBLP-C) favors higher centrality distributions to nodes from the classes having lower cardinality. To summarize, although centrality measures based on degree distribution (e.g. Degree, PageRank) are consistently biased towards the lower cardinality classes irrespective of the HIN schema followed, surprisingly Eigenvector centrality shows inconsistency in personalized schema i.e. DBLP-P. Further, it is also noted that the Closeness centrality measure is sensitive to the network schema.

Figure 7.6 presents the centrality distribution for different node classes using Personalized PageRank (PPR). As described in [105], PPR gives a personalized node ranking by considering appropriate value of damping parameter. Therefore, we consider a small damping parameter (i.e. 0.05) to assign very high weight to personalization parameter and very low weight to the network topological structure. It is evident that in the case of DBLP-C, PPR assigns higher centrality to nodes belonging to classes with lower cardinality irrespective of the personalization parameter. However, this observation is not consistent with DBLP-P where different

TABLE 7.2: Correlation of intra-class (Venue, Author, Paper) node ranking between APVTF and APV HIN variants.

Dataset	Node Class	PPR personalized by			PageRank	Degree	Betweenness	Closeness	Eigenvector
		Author	Venue	Paper					
DBLP-C	Venue	1.000	1.000	1.000	0.998	1.000	0.975	0.879	0.817
	Author	0.973	0.997	0.997	0.987	1.000	0.982	0.394	0.492
	Paper	1.000	1.000	1.000	0.986	1.000	0.368	0.002	0.111
DBLP-P	Venue	1.000	1.000	1.000	0.999	1.000	0.975	0.869	0.497
	Author	1.000	1.000	1.000	0.997	1.000	0.979	0.556	0.532
	Paper	1.000	1.000	1.000	0.994	1.000	0.400	0.256	0.356

personalization affects differently in most of the cases. Thus, it can be inferred that HINs following clique-based schema are less sensitive to the personalization parameter and more biased towards the degree distribution whereas this may not be true for personalized HIN schema.

7.6.2 Correlation Between Intra-class Ranking for APVTF and APV HINs

As observed in the above Section 7.6.1, the presence of class imbalance favors the nodes belonging to classes having low cardinality while ranking in majority of the cases. However, *Does this observation consistent for intra-class ranking when subjected to different HINs having different level of class imbalance*? To investigate this, we now estimate the Pearson Rank correlation coefficient between the intra-class ranking of APVTF and APV variants for both types of HIN schema. From Table 7.2, it is clearly visible that centrality measures namely, PPR, PageRank, and Degree have the minimum or no effect of class imbalance and yield very high correlation score. Further, for Betweenness centrality we observe a mixed result where node ranking for Venue and Author are highly correlated but node ranking for Paper class gets affected yielding a lower correlation score. However, for Closeness and Eigenvector centrality measures we observe low correlation scores for all the three intra-class node ranking. Thus, it can be inferred that intra-class node ranking is not sensitive to different levels of class imbalance for centrality measures based on degree distribution (Degree, PageRank, PPR). However, centrality measures (e.g. Betweenness, Closeness, Eigenvector) mining other properties of graphs are sensitive to class imbalance level.

7.7 Summary

This chapter studies the effect of class imbalance in mining heterogeneous bibliographic network (DBLP) using node features generated from two network embedding methods namely, DeepWalk and VERSE. To understand the effect of class imbalance, we focus on studying the performance of network embedding subjected to two different tasks namely Co-authorship prediction and Author's research area classification. For the comparative study we consider two different schemas (DBLP-C and DBLP-P) of the above discussed network and consider different variants based on the selection of node types (consequently edge types) which introduce different amount of class imbalance. We observe that reducing the class imbalance in the network does not always help in mining a heterogeneous information network (HIN) for applications of diverse nature such as link prediction, classification, etc.

As majority of the previous HIN embedding models exploit meta-paths, we further study the effect of class imbalance on meta-path-based network embedding. Although meta-path-based network embedding models are subjected to lower degree of class imbalance, we observe that they perform differently for different mining tasks. Thus, it is difficult to generalize meta-path-based network embedding for addressing class imbalance while solving different network mining problems. Further, from various experimental results, it is evident that a careful selection of important node classes helps in better embedding quality. Thus, we summarize that while mining HIN, focus should be given to selecting important node types which inherently addresses the class imbalance issue.

We further study the effects of class imbalance on inter-class and intra-class ranking using state-of-the-art centrality measures. It is evident that majority of the centrality measures are biased towards nodes belonging to lower cardinality while ranking inter-class nodes. However, the intra-class ranking is not much affected by class imbalance for random walk-based centrality measures. This work is accepted

for publication in the Journal of Informetrics[2].

[2]Anil, A., and Singh, S.R., 2020, August. "***Effect of Class Imbalance in Heterogeneous Network Embedding: An Empirical Study***". Accepted for Publication in Journal of Informetrics.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

This thesis exploits Heterogeneous Information Network (HIN) for solving link prediction problem in physical systems such as terrorist attack and bibliographic information. Unlike previous studies based on meta-paths for link prediction in HIN, we propose link prediction methods which do not require explicit meta-paths. Further, this thesis explores two themes of network science, namely (i) exploiting only the topology of the underlying network and (ii) representing given network to hidden space or exploiting network embedding to generate node features.

This thesis proposes heterogeneous transformations for traditional common neighbor-based local similarity measures and spectral graph kernels for link prediction in Chapter 3. In particular, we propose heterogeneous transformations of four popular common neighbor-based local similarity measures, namely (i) Common Neighbor, (ii) Jaccard Coefficient, (iii) Adamic-Adar Index, and (iv) Resource Allocation. Further, among spectral graph kernels we consider Path Counting, Exponential, and Neumann kernels and propose heterogeneous transformations. As different node types in HIN may have different priority/importance over different types of relations, we encode node importance for estimating likelihood of links.

Chapter 4 exploits the proposed heterogeneous similarity measures for predicting future links in a heterogeneous terrorist attack network constructed using Global Terrorism Database. This work focuses on incorporating exogenous information as node importance while predicting future links. Thus, we propose to exploit Personalized PageRank (PPR) as a single model capturing different types of exogenous information using several parametric setups. From several experimental observations, it is evident that the heterogeneous similarity measures exploiting node importance using exogenous information have positive effects on the performance of link prediction.

It is reported in earlier studies that only topological characteristics may not be enough in mining real-world networks. Further, a network latent representation may help in capturing hidden characteristics of the network. Therefore, Chapter 5 studies the network embedding models for generating node features which can be further exploited for various network-mining tasks such as node classification, link prediction, clustering, etc. In particular, we focus on recently proposed neural network-based network embedding methods for HIN embedding. As majority of the HIN embedding methods exploit meta-path, this chapter critically analyzes the applicability of different meta-paths for HIN embedding. It is observed that meta-path-based HIN embedding may not be efficient as it loses network information. Moreover, meta-paths are task sensitive and the embeddings cannot be generalized.

Selecting optimal meta-paths is not a trivial task. Therefore, Chapter 6 proposes a novel network embedding method for HINs that does not require explicit meta-paths. We exploit $\boldsymbol{k}$-hop random walks to generate node sequences which are further used to train the skip-gram-based neural network model. It is observed from various experimental results that the proposed network embedding method outperforms state-of-the-art baselines for link prediction task in majority of the cases.

Since HIN construction permits irregular number of instances for various types of nodes and relations, class imbalance seems to be an inherent characteristics of HIN. Therefore, Chapter 7 studies the effect of class imbalance in HIN on

network embedding for solving tasks of diverse nature, i.e. link prediction and node classification. It is observed that class imbalance affects HIN embedding differently over different tasks. Further, selecting node types is as critical as addressing class imbalance for an efficient HIN embedding.

8.2 Limitation

This thesis work mainly focuses on predicting future links in HIN. Thus, it is limited to exploring state-of-the-art methods which are generic and can be applied over different HINs. Consequently, this thesis do not present literature for previous link prediction methods proposed for mining specific types of HINs. As estimating best meta-path is not trivial, the previous studies for link prediction using meta-path is beyond the scope of this thesis.

8.3 Future Works

As noted from the above studies, different HIN schema for the same system has different effects over solving various tasks. Further, there is no standard protocols to construct HINs for real-world systems. Thus, studying and setting standards for HIN construction is a potential future direction.

Majority of the studies on network embedding in HIN do not consider temporal characteristics of HIN. However, as observed in Chapter 4 that incorporating temporal dynamics helps in boosting performance for link prediction task. Thus, HIN embedding exploiting temporal information may be a future extension.

Bibliography

[1] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable representation learning for heterogeneous networks," in *23rd Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 135–144.

[2] J. Han, "Mining heterogeneous information networks by exploring the power of links," in *12th Proceedings of the International Conference on Discovery Science*, 2009, pp. 13–30.

[3] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[4] Y. Sun and J. Han, "Mining heterogeneous information networks: a structural analysis approach," *Acm Sigkdd Explorations Newsletter*, vol. 14, no. 2, pp. 20–28, 2013.

[5] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *18th Proceedings of the International Conference on Machine Learning*, 2001, pp. 282–289.

[6] K. Wakita and T. Tsurumi, "Finding community structure in mega-scale social networks," in *16th Proceedings of the International Conference on World Wide Web*, 2007, pp. 1275–1276.

[7] M. Sales-Pardo, R. Guimera, A. A. Moreira, and L. A. N. Amaral, "Extracting the hierarchical organization of complex systems," *National Academy of Sciences*, vol. 104, no. 39, pp. 15 224–15 229, 2007.

[8] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *12th Proceedings of the International Conference on Information and Knowledge Management*, 2003, pp. 556–559.

[9] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *7th Proceedings of International Conference on World-Wide Web*, 1998, pp. 107–117.

[10] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[11] H. Ma, I. King, and M. R. Lyu, "Learning to recommend with social trust ensemble," in *32nd Proceedings of the International Conference on Research and Development in Information Retrieval*, 2009, pp. 203–210.

[12] X. Yang, H. Steck, and Y. Liu, "Circle-based recommendation in online social networks," in *18th Proceedings of the International Conference on Knowledge Discovery and Data mining*, 2012, pp. 1267–1275.

[13] C. Shi, Y. Li, J. Zhang, Y. Sun, and S. Y. Philip, "A survey of heterogeneous information network analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 17–37, 2017.

[14] X. Liu, Y. Yu, C. Guo, and Y. Sun, "Meta-path-based ranking with pseudo relevance feedback on heterogeneous graph for citation recommendation," in *23rd Proceedings of the International Conference on Information and Knowledge Management*, 2014, pp. 121–130.

[15] Y. Li, C. Shi, P. S. Yu, and Q. Chen, "Hrank: A path based ranking method in heterogeneous information network," in *15th Proceedings of the Web-Age Information Management*, 2014, pp. 553–565.

[16] S. Lee, S. Park, M. Kahng, and S.-g. Lee, "Pathrank: A novel node ranking measure on a heterogeneous graph for recommender systems," in *21st Proceedings of the International Conference on Information and Knowledge Management*, 2012, pp. 1637–1641.

[17] M.-H. Tsai, C. Aggarwal, and T. Huang, "Ranking in heterogeneous social media," in *7th Proceedings of the International Conference on Web Search and Data Mining*, 2014, pp. 613–622.

[18] M. Ji, J. Han, and M. Danilevsky, "Ranking-based classification of heterogeneous information networks," in *17th Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 1298–1306.

[19] C. Shi, C. Zhou, X. Kong, P. S. Yu, G. Liu, and B. Wang, "Heterecom: A semantic-based recommendation system in heterogeneous networks," in *18th Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 1552–1555.

[20] X. Kong, P. S. Yu, Y. Ding, and D. J. Wild, "Meta path-based collective classification in heterogeneous information networks," in *21st Proceedings of the International Conference on Information and knowledge Management*, 2012, pp. 1567–1571.

[21] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, "Co-author relationship prediction in heterogeneous bibliographic networks," in *3rd Proceedings of International Conference on Advances in Social Networks Analysis and Mining*, 2011, pp. 121–128.

[22] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla, "When will it happen?: relationship prediction in heterogeneous information networks," in *5th Proceedings of the International Conference on Web Search and Data Mining*, 2012, pp. 663–672.

[23] X. Yu, Q. Gu, M. Zhou, and J. Han, "Citation prediction in heterogeneous bibliographic networks," in *12th Proceedings of the SIAM International Conference on Data Mining*, 2012, pp. 1119–1130.

[24] J. Chen, H. Gao, Z. Wu, and D. Li, "Tag co-occurrence relationship prediction in heterogeneous information networks," in *19th International Conference on Parallel and Distributed Systems*, 2013, pp. 528–533.

[25] B. Cao, X. Kong, and S. Y. Philip, "Collective prediction of multiple types of links in heterogeneous information networks," in *14th Proceedings of International Conference on Data Mining*, 2014, pp. 50–59.

[26] J. Kim and T. Wilhelm, "What is a complex graph?" *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 11, pp. 2637–2652, 2008.

[27] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.

[28] M. E. Shaw, "Group structure and the behavior of individuals in small groups," *The Journal of psychology*, vol. 38, no. 1, pp. 139–149, 1954.

[29] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[30] A. Bavelas, "Communication patterns in task-oriented groups," *The Journal of the Acoustical Society of America*, vol. 22, no. 6, pp. 725–730, 1950.

[31] P. Bonacich, "Power and centrality: A family of measures," *American journal of sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[32] G. Salton, "Automatic text processing: The transformation, analysis, and retrieval of reading," *Addison-Wesley*, vol. 169, 1989.

[33] H. Chen, X. Li, and Z. Huang, "Link prediction approach to collaborative filtering," in *5th Proceedings of the Joint Conference on Digital Libraries*, 2005, pp. 141–142.

[34] Z. Huang and D. K. Lin, "The time-series link prediction problem with applications in communication surveillance," *INFORMS Journal on Computing*, vol. 21, no. 2, pp. 286–303, 2009.

[35] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bulletin del la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.

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

[37] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *The European Physical Journal B*, vol. 71, no. 4, pp. 623–630, 2009.

[38] B. Nettasinghe and V. Krishnamurthy, "What do your friends think? efficient polling methods for networks using friendship paradox," *arXiv preprint arXiv:1802.06505*, 2018.

[39] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *19th Proceedings of the International Conference on World Wide Web*, 2010, pp. 851–860.

[40] L. Seeman and Y. Singer, "Adaptive seeding in social networks," in *54th Annual Symposium on Foundations of Computer Science*, 2013, pp. 459–468.

[41] B. Nettasinghe and V. Krishnamurthy, "Influence maximization over markovian graphs: A stochastic optimization approach," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 5, no. 1, pp. 1–14, 2019.

[42] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *9th Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2003, pp. 137–146.

[43] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[44] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[45] J. Jiang, "Rumor source identification in complex networks," Deakin Univeristy, Tech. Rep., 2017.

[46] M. Ostilli, E. Yoneki, I. X. Leung, J. F. Mendes, P. Lió, and J. Crowcroft, "Statistical mechanics of rumour spreading in network communities," *Procedia Computer Science*, vol. 1, no. 1, pp. 2331–2339, 2010.

[47] M. T. Schaub, J.-C. Delvenne, M. Rosvall, and R. Lambiotte, "The many facets of community detection in complex networks," *Applied network science*, vol. 2, no. 1, p. 4, 2017.

[48] D. Yu, M. Righero, and L. Kocarev, "Estimating topology of networks," *Physical Review Letters*, vol. 97, no. 18, p. 188701, 2006.

[49] M. Timme, "Revealing network connectivity from response dynamics," *Physical review letters*, vol. 98, no. 22, p. 224101, 2007.

[50] D. Zhou, Y. Xiao, Y. Zhang, Z. Xu, and D. Cai, "Causal and structural connectivity of pulse-coupled nonlinear networks," *Physical review letters*, vol. 111, no. 5, p. 054102, 2013.

[51] M. Timme and J. Casadiego, "Revealing networks from dynamics: an introduction," *Journal of Physics A: Mathematical and Theoretical*, vol. 47, no. 34, p. 343001, 2014.

[52] M. Nitzan, J. Casadiego, and M. Timme, "Revealing physical interaction networks from statistics of collective dynamics," *Science advances*, vol. 3, no. 2, p. e1600396, 2017.

[53] X. Li and X. Li, "Reconstruction of stochastic temporal networks through diffusive arrival times," *Nature communications*, vol. 8, p. 15729, 2017.

[54] C. Ma, H.-F. Zhang, and Y.-C. Lai, "Reconstructing complex networks without time series," *Physical Review E*, vol. 96, no. 2, p. 022320, 2017.

[55] R.-Q. Su, Y.-C. Lai, and X. Wang, "Identifying chaotic fitzhugh–nagumo neurons using compressive sensing," *Entropy*, vol. 16, no. 7, pp. 3889–3902, 2014.

[56] Z. Shen, W.-X. Wang, Y. Fan, Z. Di, and Y.-C. Lai, "Reconstructing propagation networks with natural diversity and identifying hidden sources," *Nature communications*, vol. 5, p. 4323, 2014.

[57] G. Mei, X. Wu, Y. Wang, M. Hu, J.-A. Lu, and G. Chen, "Compressive-sensing-based structure identification for multilayer networks," *IEEE transactions on cybernetics*, vol. 48, no. 2, pp. 754–764, 2018.

[58] W.-X. Wang, R. Yang, Y.-C. Lai, V. Kovanis, and M. A. F. Harrison, "Time-series–based prediction of complex oscillator networks via compressive sensing," *EPL (Europhysics Letters)*, vol. 94, no. 4, p. 48006, 2011.

[59] N. Sett, "Exploiting tie-strength and structure towards link prediction in social networks," Ph.D. dissertation, 2017.

[60] J. Tague-Sutcliffe, "An introduction to informetrics," *Information processing & management*, vol. 28, no. 1, pp. 1–3, 1992.

[61] B. A. Huberman, *The Laws of the Web*. MIT Press, 2001.

[62] L. A. Adamic, R. M. Lukose, and B. A. Huberman, "13 local search in unstructured networks," *Handbook of graphs and networks*, p. 295, 2003.

[63] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman, "Search in power-law networks," *Physical review E*, vol. 64, no. 4, p. 046135, 2001.

[64] A. Iamnitchi, M. Ripeanu, and I. Foster, "Locating data in (small-world?) peer-to-peer scientific collaborations," in *1st Proceedings of the International Workshop on Peer-to-Peer Systems*, 2002, pp. 232–241.

[65] D. E. Knuth, *The Stanford GraphBase: a platform for combinatorial computing*. AcM Press New York, 1993.

[66] O. Kinouchi, A. S. Martinez, G. F. Lima, G. M. Lourenço, and S. Risau-Gusman, "Deterministic walks in random networks: An application to thesaurus graphs," *Physica A: Statistical Mechanics and its Applications*, vol. 315, no. 3-4, pp. 665–676, 2002.

[67] M. Steyvers and J. B. Tenenbaum, "The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth," *Cognitive science*, vol. 29, no. 1, pp. 41–78, 2005.

[68] A. E. Motter, A. P. De Moura, Y.-C. Lai, and P. Dasgupta, "Topology of the conceptual network of language," *Physical Review E*, vol. 65, no. 6, p. 065102, 2002.

[69] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no. 12, pp. 61–71, 1992.

[70] P. Resnick and H. R. Varian, "Recommender systems," *Communications of the ACM*, vol. 40, no. 3, pp. 56–59, 1997.

[71] U. Shardanand and P. Maes, "Social information filtering: Algorithms for automating word of mouth," in *13th Proceedings of the Conference on Human Factors in Computing Systems*, 1995, pp. 210–217.

[72] Y. Sun, Y. Yu, and J. Han, "Ranking-based clustering of heterogeneous information networks with star network schema," in *15th Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 797–806.

[73] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," *VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2011.

[74] Q. Ou, Y.-D. Jin, T. Zhou, B.-H. Wang, and B.-Q. Yin, "Power-law strength-degree correlation from resource-allocation dynamics on weighted networks," *Physical Review E*, vol. 75, no. 2, pp. 021 102–1–021 102–5, 2007.

[75] F. Fouss, L. Yen, A. Pirotte, and M. Saerens, "An experimental investigation of graph kernels on a collaborative recommendation task," in *6th International Conference on Data Mining*, 2006, pp. 863–868.

[76] F. Fouss, K. Francoisse, L. Yen, A. Pirotte, and M. Saerens, "An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification," *Neural networks*, vol. 31, pp. 53–72, 2012.

[77] C. Wang, V. Satuluri, and S. Parthasarathy, "Local probabilistic models for link prediction," in *7th Proceedings of International Conference on Data Mining*, 2007, pp. 322–331.

[78] D. Billsus and M. J. Pazzani, "Learning collaborative information filters." in *15th Proceedings of the International Conference on Machine Learning*, vol. 98, 1998, pp. 46–54.

[79] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Application of dimensionality reduction in recommender system-a case study," DTIC Document, Tech. Rep., 2000.

[80] A. K. Menon and C. Elkan, "Link prediction via matrix factorization," in *11th Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2011, pp. 437–452.

[81] J. Kunegis, D. Fay, and C. Bauckhage, "Network growth and the spectral evolution model," in *19th Proceedings of the International Conference on Information and Knowledge Management*, 2010, pp. 739–748.

[82] T. Thorne and M. P. Stumpf, "Graph spectral analysis of protein interaction network evolution," *Journal of The Royal Society Interface*, vol. 9, no. 75, pp. 2653–2666, 2012.

[83] S. Sarkar and A. Dong, "Community detection in graphs using singular value decomposition," *Physical Review E*, vol. 83, no. 4, p. 046114, 2011.

[84] F. R. Chung, *Spectral graph theory*. American Mathematical Soc., 1997, vol. 92.

[85] D. M. Cvetković, P. Rowlinson, and S. Simic, *Eigenspaces of graphs*. Cambridge University Press, 1997, no. 66.

[86] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Physical Review E*, vol. 80, no. 4, p. 046122, 2009.

[87] R. I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete input spaces," in *19th Proceedings of the International Conference on Machine Learning*, vol. 2, 2002, pp. 315–322.

[88] J. Kunegis and A. Lommatzsch, "Learning spectral graph transformations for link prediction," in *26th Proceedings of the International Conference on Machine Learning*, 2009, pp. 561–568.

[89] H. Cai, V. W. Zheng, and K. Chang, "A comprehensive survey of graph embedding: problems, techniques and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 9, pp. 1616–1637, 2018.

[90] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *15th Proceedings of the Advances in neural information processing systems*, 2002, pp. 585–591.

[91] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[92] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[93] A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola, "Distributed large-scale natural graph factorization," in *22nd Proceedings of the International Conference on World Wide Web*, 2013, pp. 37–48.

[94] M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu, "Asymmetric transitivity preserving graph embedding," in *22nd Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1105–1114.

[95] S. Cao, W. Lu, and Q. Xu, "Grarep: Learning graph representations with global structural information," in *24th Proceedings of the International Conference on Information and Knowledge Management*, 2015, pp. 891–900.

[96] ——, "Deep neural networks for learning graph representations." in *30th Proceedings of the AAAI Conference on Artificial Intelligence*, 2016, pp. 1145–1152.

[97] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," in *22nd Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 855–864.

[98] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *20th Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 701–710.

[99] C. Tu, W. Zhang, Z. Liu, and M. Sun, "Max-margin deepwalk: Discriminative learning of network representation." in *25th Proceedings of the International Joint Conference on Artificial Intelligence*, 2016, pp. 3889–3895.

[100] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *22nd Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1225–1234.

[101] L. F. Ribeiro, P. H. Saverese, and D. R. Figueiredo, "struc2vec: Learning node representations from structural identity," in *23rd Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 385–394.

[102] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," in *24th Proceedings of the International Conference on World Wide Web*, 2015, pp. 1067–1077.

[103] A. Tsitsulin, D. Mottin, P. Karras, and E. Müller, "Verse: Versatile graph embeddings from similarity measures," in *27th Proceedings of the International Conference on World Wide Web*, 2018, pp. 539–548.

[104] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *26th Proceedings of the Advances in neural information processing systems*, 2013, pp. 3111–3119.

[105] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," in *7th Proceedings of the International Conference on World Wide Web*, 1998, pp. 161–172.

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

[107] J. A. Rodrguez and J. A. Rodrguez, "The march 11 th terrorist network: In its weakness lies its strength," 2005.

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

[109] "National consortium for the study of terrorism and responses to terrorism (start). global terrorism database [data file]." [Online]. Available: https://www.start.umd.edu/

[110] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[111] R. Lichtnwalter and N. V. Chawla, "Link prediction: fair and effective evaluation," in *4th Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 2012, pp. 376–383.

[112] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Physical Review E*, vol. 80, no. 4, p. 046122, 2009.

[113] N. Sett, S. R. Singh, and S. Nandi, "Influence of edge weight on node proximity based link prediction methods: An empirical analysis," *Neurocomputing*, vol. 172, pp. 71–83, 2016.

[114] Y. Dong, J. Tang, S. Wu, J. Tian, N. V. Chawla, J. Rao, and H. Cao, "Link prediction and recommendation across heterogeneous social networks," in *12th Proceedings of the International Conference on Data Mining*, 2012, pp. 181–190.

[115] B. M. Jenkins, "The new age of terrorism," *The McGraw-Hill homeland security handbook*, pp. 117–130, 2006.

[116] B. L. Nacos, "Terrorism/counterterrorism and media in the age of global communication," in *United Nations University Global Seminar Second Shimame-Yamaguchi Session, Terrorism-A Global Challenge*, vol. 5, 2006.

[117] S. A. Myers, C. Zhu, and J. Leskovec, "Information diffusion and external influence in networks," in *18th Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 33–41.

[118] T. H. Haveliwala, "Topic-sensitive pagerank," in *11th Proceedings of the International Conference on World Wide Web*, 2002, pp. 517–526.

[119] W. Liu and L. Lü, "Link prediction based on local random walk," *EPL (Europhysics Letters)*, vol. 89, no. 5, p. 58007, 2010.

[120] K. Pearson, "The problem of the random walk," *Nature*, vol. 72, no. 1867, p. 342, 1905.

[121] D. M. Dunlavy, T. G. Kolda, and E. Acar, "Temporal link prediction using matrix and tensor factorizations," *Transactions on Knowledge Discovery from Data*, vol. 5, no. 2, p. 10, 2011.

[122] R. K. Pan and J. Saramäki, "Path lengths, correlations, and centrality in temporal networks," *Physical Review E*, vol. 84, no. 1, p. 016105, 2011.

[123] Y. Sun and J. Han, "Mining heterogeneous information networks: principles and methodologies," *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 3, no. 2, pp. 1–159, 2012.

[124] T.-y. Fu, W.-C. Lee, and Z. Lei, "Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning," in *27th Proceedings of the Conference on Information and Knowledge Management*, 2017, pp. 1797–1806.

[125] D. Yang, Y. Xiao, B. Xu, H. Tong, W. Wang, and S. Huang, "Which topic will you follow?" in *12th Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2012, pp. 597–612.

[126] Y. Wu, X. Zhang, Y. Bian, Z. Cai, X. Lian, X. Liao, and F. Zhao, "Second-order random walk-based proximity measures in graph analysis: formulations and algorithms," *The International Journal on Very Large Data Bases*, vol. 27, no. 1, pp. 127–152, 2018.

[127] J. Tang, M. Qu, and Q. Mei, "Pte: Predictive text embedding through large-scale heterogeneous text networks," in *21st Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1165–1174.

[128] R. Hussein, D. Yang, and P. Cudré-Mauroux, "Are meta-paths necessary?: Revisiting heterogeneous graph embeddings," in *27th Proceedings of the International Conference on Information and Knowledge Management*, 2018, pp. 437–446.

[129] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imbalanced data sets," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 1–6, 2004.

[130] A. Anil, U. Chugh, and S. R. Singh, "On applying meta-path for network embedding in mining heterogeneous DBLP network," *CoRR*, vol. abs/1808.04799, 2018.

[131] Z. Huang and N. Mamoulis, "Heterogeneous information network embedding for meta path based proximity," *CoRR*, vol. abs/1701.05291, 2017.

[132] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu, "Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks," *Transactions on Knowledge Discovery from Data*, vol. 7, no. 3, p. 11, 2013.

[133] Y. Dong, J. Zhang, J. Tang, N. V. Chawla, and B. Wang, "Coupledlp: Link prediction in coupled networks," in *21st Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 199–208.

[134] C. Shi, B. Hu, W. X. Zhao, and S. Y. Philip, "Heterogeneous information network embedding for recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 2, pp. 357–370, 2019.

[135] J. Shang, M. Qu, J. Liu, L. M. Kaplan, J. Han, and J. Peng, "Meta-path guided embedding for similarity search in large-scale heterogeneous information networks," *arXiv preprint arXiv:1610.09769*, 2016.

[136] T. Chen and Y. Sun, "Task-guided and path-augmented heterogeneous network embedding for author identification," in *10th Proceedings of the International Conference on Web Search and Data Mining*, 2017, pp. 295–304.

[137] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 148–156.

[138] Y. Liu, N. Chawla, E. Shriberg, A. Stolcke, and M. Harper, "Resampling techniques for sentence boundary detection: a case study in machine learning from imbalanced data for spoken language processing," *Computer Speech and Language, to appear*, 2003.

[139] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. of the Intl Conf. on Artificial Intelligence*, 2000.

[140] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[141] C. Wang, Y. Sun, Y. Song, J. Han, Y. Song, L. Wang, and M. Zhang, "Relsim: relation similarity search in schema-rich heterogeneous information networks," in *16th Proceedings of the SIAM International Conference on Data Mining*, 2016, pp. 621–629.

[142] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in *22nd Proceedings of the International Conference on Information & Knowledge Management*, 2013, pp. 179–188.

[143] R. Angelova, G. Kasneci, and G. Weikum, "Graffiti: graph-based classification in heterogeneous networks," *World Wide Web*, vol. 15, no. 2, pp. 139–170, 2012.

[144] C. Shi, X. Kong, P. S. Yu, S. Xie, and B. Wu, "Relevance search in heterogeneous networks," in *15th Proceedings of the International Conference on Extending Database Technology*, 2012, pp. 180–191.

[145] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu, "Rankclus: integrating clustering with ranking for heterogeneous information network analysis," in *12th Proceedings of the International Conference on Extending Database Technology: Advances in Database Technology*, 2009, pp. 565–576.

[146] J. Kuck, H. Zhuang, X. Yan, H. Cam, and J. Han, "Query-based outlier detection in heterogeneous information networks," in *18th Proceedings of the International Conference on Extending Database Technology*, 2015, pp. 325–336.

[147] E. Zhong, W. Fan, J. Wang, L. Xiao, and Y. Li, "Comsoc: adaptive transfer of user behaviors over composite social network," in *18th Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 696–704.

[148] H. Huang, A. Zubiaga, H. Ji, H. Deng, D. Wang, H. Le, T. Abdelzaher, J. Han, A. Leung, J. Hancock *et al.*, "Tweet ranking based on heterogeneous networks," *24th Proceedings of the International Conference on Computational Linguistics*, pp. 1239–1256, 2012.

[149] J. D. Cruz, C. Bothorel, and F. Poulet, "Integrating heterogeneous information within a social network for detecting communities," in *5th Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 2013, pp. 1453–1454.

[150] A. Anil, D. Kumar, S. Sharma, R. Singha, R. Sarmah, N. Bhattacharya, and S. R. Singh, "Link prediction using social network analysis over heterogeneous terrorist network," in *8th Proceedings of the International Conference on SocialCom*, 2015, pp. 267–272.

[151] A. Anil, S. R. Singh, and R. Sarmah, "Personalised pagerank as a method of exploiting heterogeneous network for counter terrorism and homeland security," in *15th Proceedings of the International Conference on Web Intelligence*, 2016, pp. 327–334.

[152] A. Anil, S. Singhal, P. Jain, S. R. Singh, A. Ladhar, S. Singh, and U. Chugh, "Network sampling using k-hop random walks for heterogeneous network embedding," in *6th Proceedings of the Joint International Conference on Data Science and Management of Data*, 2019, pp. 354–357.

Publications (Related to Thesis)

Conference:

- **Published**

1. **Anil, A.**, Kumar, D., Sharma, S., Singha, R., Sarmah, R., Bhattacharya, N. and Singh, S.R., 2015, December. "*Link prediction using social network analysis over heterogeneous terrorist network*". In 8th Proceedings of the International Conference on SocialCom (pp. 267-272). IEEE, Chengdu, China.
2. **Anil, A.**, Singh, S.R. and Sarmah, R., 2016, October. "*Personalised PageRank as a Method of Exploiting Heterogeneous Network for Counter Terrorism and Homeland Security*". In 15th Proceedings of International Conference on Web Intelligence (pp. 327-334). IEEE, Omaha, Nebraska, USA.
3. **Anil, A.**, Singhal, S., Jain, P., Singh, S.R., Ladhar, A., Singh, S. and Chugh, U., 2019, January. "*Network Sampling Using k-hop Random Walks for Heterogeneous Network Embedding*". In 6th Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (pp. 354-357). ACM, Kolkata, India.
4. **Anil, A.**, Chugh, U., and Singh, S. R. 2019, December. "*On Applying Meta-path for Network Embedding in Mining Heterogeneous DBLP Network*". In 8th Proceedings of the International Conference on Pattern Recognition and Machine Intelligence (pp. 249-257). Springer, Tezpur, India.

- **Accepted**

1. **Anil, A.**, Chugh, U. and Singh, S.R., 2018, August. "*On Applying Meta-path for Network Embedding in Mining Heterogeneous DBLP Network*". International Conference on Asia-Pacific Digital Libraries (ICADL), Hamilton, New Zealand.

2. **Anil, A.**, Ladhar, A., Singh, S. and Chugh, U., Singh, S.R., 2018, August. "*Network Sampling Using k-hop Random Walks for Heterogeneous Network Embedding*". International Conference on Asia-Pacific Digital Libraries (ICADL), Hamilton, New Zealand.

Journal:

- **Published**

1. **Anil, A.**, Singh, S.R. and Sarmah, R., 2018, January. "*Mining heterogeneous terrorist attack network using personalized PageRank*". In Web Intelligence (Vol. 16, No. 1, pp. 37-52)

- **Accepted**

1. **Anil, A.**, and Singh, S.R., 2020, January. "*Effect of Class Imbalance in Heterogeneous Network Embedding: An Empirical Study*". Journal of Informetrics.

Other Publications

Conference:

- **Anil, A.**, Sett, N., and Singh, S.R. , 2014, July. "*Modeling evolution of a social network using temporal graph kernels*". In 37th Proceedings of international ACM SIGIR conference on Research and development in information retrieval (pp. 1051-1054). ACM, Gold Coast, Australia.

Journal:

- Singh, L.G., **Anil, A.**, and Singh, S.R., 2020, January. "*SHE: Sentiment Hashtag Embedding Through Multitask learning*". In IEEE Transactions on Computational Social Systems (Accepted).

Brief Biography of the Author

Akash Anil was born in Jamshedpur, Jharkhand, India on 26th January 1987. After completing his basic education from Jamshedpur, he completed Bachelor of Technology (B.Tech) in Dept. of Computer Science & Engineering from Synergy Institute of Engineering & Technology, Dhenkanal, Orissa in the year 2009. After graduation he served as a lecturer in the Dept. of Information Technology at Gyan Ganga Institute of Technology & Management, Bhopal for two and a half years. He competed Master of Technology (M.Tech) degree in the Dept. of Computer Science & Engineering from Indian Institute of Technology Guwahati in the year 2014. Thereafter, he was enrolled as a Ph.D. research scholar in the Dept. of Computer Science & Engineering at Indian Institute of Technology Guwahati. In Ph.D., he was supervised by Dr. Sanasam Ranbir Singh. His research interests lie in the area of Social Network Analysis, Machine Learning, and Deep Learning on Graph.