國立臺灣大學電機資訊學院資訊工程研究所

碩士論文

Graduate Institute of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

異質性社群網路之擴增性意涵導向個體檢索

Semantic-Driven Explorative Entity Search in Heterogeneous Social Networks

龔鵬驊

Perng-Hwa Kung

指導教授：林守德 博士

Advisor: Shou-De Lin, Ph.D.

中華民國 102 年 7 月

July, 2013

Acknowledgement

I would like to thank my advisor, Professor Shou-De Lin, for his kind introduction to the realm of social network analysis and machine learning. The past two years have truly been quite a ride. Professor Lin introduced me to many different aspects of computer science research, and I learnt a great deal of proper research methodology in the process. I really appreciate Professor's patience and tolerance to my relatively incompetent self, both in terms of research and work. I would like to also thank all the instructors and Professors for giving invaluable knowledge.

I would also like to express my gratitude to the seniors in the lab (those who are still in the lab and those who are already graduates) for sharing valuable experiences and tips to working in teams to tackle challenging research problems. I specifically want to thank my labmates for their kindness and making me feel comfortable at the lab.

In particular, my deepest gratitude goes to my dear friends, who are always there for support. Without them, I wouldn't have endured all the hardships. In life, some good old pals are always the best treasures out there. Thank you, Ben, Jim, Daniel, Ken, Steve, Kang, and other good friends along the way.

Last but not least, my deepest thanks go to my family, who are always supportive of my choice. This has not been an easy journey, but I have pulled through. Thank you and I love you, Mom and Dad.

And yes, I would like to express my greatest appreciation to my girlfriend, Ying-Ju, for the past year for her unreserved love and support. I love you.

Abstract

Heterogeneous social network has seen a rapid rise in research and industry interest in the widely popularizing online social or information networks, such as Twitter, Facebook, or Google Knowledge Graph. Such networks are characterized by large-scale of data volume, and the varying multitude of roles that an individual (or entity) plays and interacts with other members of the network. Oftentimes, engineers that design applications to exploit the wealth of information hidden within the networks need to extract parts of the network and semantically similar entities to the target of interest via techniques such as crawling. This process faces the challenge that one very frequently does not have the permission to access the fully observed graph for network services at large.

This study defines and examines the problem of exploratively searching semantically related nodes in heterogeneous social network, under the context of specific meta-path semantics dictated by the graph schema of the network. In particular, the paper proposes a framework to sequentially crawl entities from the full network, where at each stage, the process calculates the expected scores for the candidate nodes using metrics that measure meta-path similarity. Moreover, we propose score propagation heuristics to facilitate the estimation of such expected scores. Experiments on several real world networks reveal that the proposed methods can estimate meta-path semantic metrics using little exploration costs across various meta-path semantics. In addition, effects on different parameter settings are tested. Lastly, the study explores applying sampled nodes to reflect ability to identify group membership and train ranking models on attributes alone for metric score prediction.

Keywords: Social Network Analysis, Heterogeneous Network, Entity Search, Sample by Exploration

摘要

由於網際網路的社群網路服務崛起，異質性社群網路分析近年來在產業界及學界得到高度的關注。此類網路涵括Twitter, Facebook, 甚或是資料網路如Google的Knowledge Graph等。這些社群網路通常具有資料量大、及網路中個體有獨特的身分或特殊的互動狀況等特性。亦即，網路中的點(個體)及邊(互動狀況)代表特定的種類。在設計特定的應用中，要有效地從這些資料中挖掘出特定的隱藏資訊，工程師通常需要利用像網路爬蟲的技巧一筆一筆從伺服器中讀取資訊。此類型的應用的一大挑戰在於，使用者通常無法直接擷取所有網路的相關資訊。在受權限限制的環境下，如何有效在最小成本中抓取型態相似的個體成為非常重要的課題。

本篇研究主要探討並定義在異質性網路中的擴增性意涵導向個體檢索：給定意涵路徑(meta-path)跟量度方式，將一步步從完整的網路中抓取相似的個體。此研究提出一個一般化的解決方案，在每個取樣的步驟時，賦予感興趣的點一個意涵路徑相似度的期望值，並進行分數權衡過的取樣。此外，將提出 MetaRank，一個網頁排名的變形，進行期望值的估算。實驗結果於數個現實生活中的社群網路顯示提出的方法可有效在有限的搜索成本中估算意涵路徑度量的相似度。實驗亦將探討不同參數的設定。最後，使用附屬屬性來輔助預測模型的建立及擴增性檢索的準確度驗證此方法的延伸性。

關鍵字：社群網路分析，異質性網路，個體檢索，擴增性取樣

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The widespread hype in online social networks sees an ever-increasing load of data that presents practitioners opportunities to tap unlimited potential of deep knowledge. However, these new sources of data are no petty beasts to tame: these networks are often heterogeneous and dictated by complex semantic structures, which are mainly classified as interrelations between individual attributes and inter-person interactions. In fact, many notable online social network services can be characterized as heterogeneous social networks. Some real-life evidences are exemplary: activity relationships in Facebook [2]; different interaction behaviors between users in Youtube [15]; and different group memberships across communities in Twitter[8]. Take publication network DBLP[1] as a clarifying example, where we provide the graph schema in Figure 1, entities can be authors, papers, etc. Interactions can occur through an author writing a paper, a paper citing another paper, etc. The concept of heterogeneous type information for nodes and edges in a network adds a new dimension to network topology, and has attracted high interest for research and industry practice.

To date, typical network analysis relies on topological properties of the graph itself. Many works have emphasized on various aspects of graph properties, especially on how topological information can be used to cluster communities[5], predict relationships[10],

[1] http://www.informatik.uni-trier.de/~ley/db/

determine roles of an individual[13], etc. However, until very recently, most works stressed little on incorporating structured type information into feature generation or data modeling. More popular usage of node type information was on population estimation[3] or regarded as only context information, rather than exploring the connectivity across graphs between subgraphs of different node types. To go beyond the traditional realm of graph analysis, heterogeneous network analysis taps topological interactions between nodes of different types by noting that homogeneous network is just a projection on information plane from heterogeneous network[2]. A flurry of works on heterogeneous network discusses on issues such as link prediction, clustering, network sampling, and similarity search. We take particular interest on the topic of similarity entity search within heterogeneous network and the rest of the paper will discuss solely on this specific subtopic.

A particular problem relevant to heterogeneous network is node (entity) similarity search. In this problem, we are interested in determining how closely related are the nodes of particular types (e.g. the author similar to a particular scholar, or the conference that scholar is most likely to take interest in). Generally, searching in heterogeneous network provides a merit that users can incorporate semantic information directly into the crawler module or the like in their applications. In order to search semantically, we need a semantic structure and corresponding metrics to capture the similarity between two entity objects. For the generic semantic structure used in the search task, we apply the recently proposed meta-path semantics[14]. A meta-path is a sequence of node types $P = (T_1,\ldots T_N)$, denoting how the two entities of types at the ends of the sequence (i.e. T_1 and T_N) are semantically connected in a network. Take Figure 1 as example, to relate two authors, the paper identifies a common venue where both

[2] KDD 2012 Summer School, Jiawei Han: http://kdd2012.sigkdd.org/summer_school.shtml#han

authors publish paper in. Consider movie network as a different example, where a movie may involve actors, directors, production firms, and other various attributes, to compare the similarity between two actors, we may be interested in the common collaborations with particular director (Actor–Movie–Director-Movie-Actor sequence), and this information is often not easily recognizable by movie titles alone. Obviously, meta-path presents flexibility towards semantic interpretation between entities.

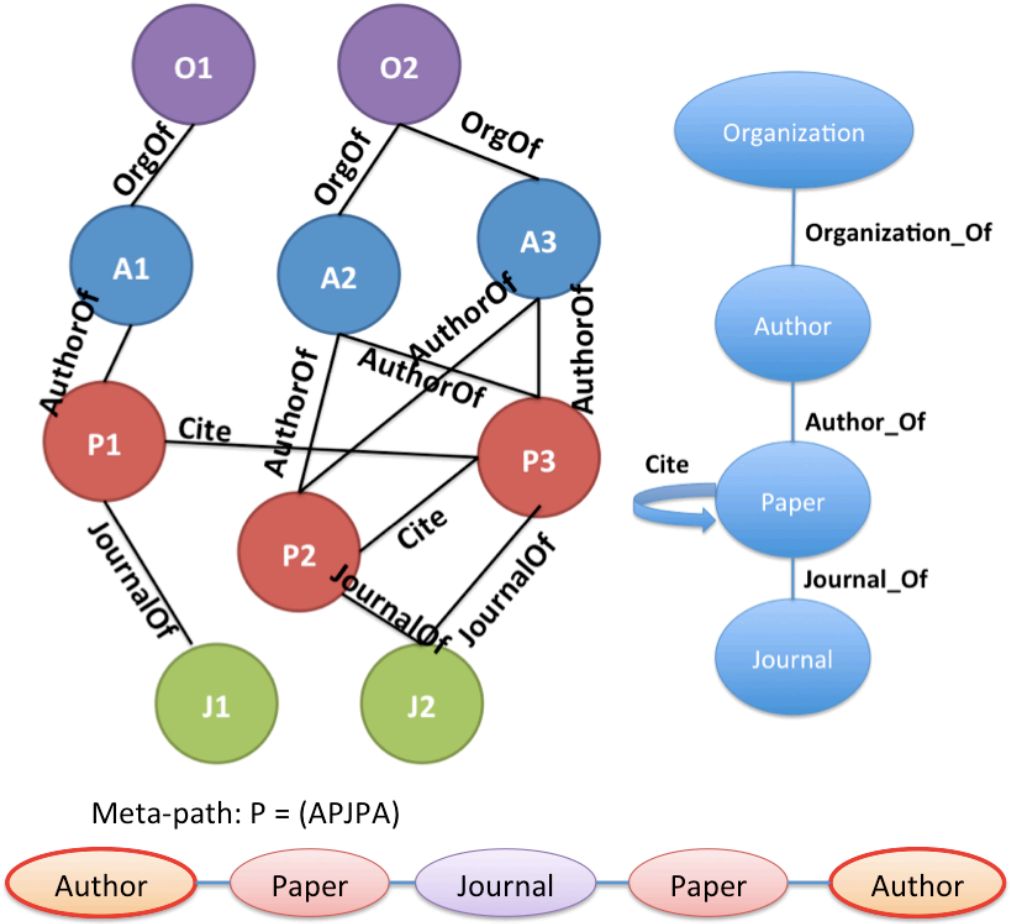


Figure 1-1. Above Right: An illustration of publication network; Above Left: Corresponding graph schema for different node and link types; Below: A meta-path semantic relating two objects of node type "Author".

However, existing works on similarity search in heterogeneous network mainly focus on server-side applications: that is, the information of entire network is available beforehand. This assumption does not apply to user-side scenarios, as one immediate

problem is the limited accessibility of the network. Many online social networks such as Facebook do not reveal the complete observed graph, thus making approaches that require access of the full network infeasible (such as direct full matrix computation). Furthermore, as we need to inquire the server every time the crawler searches, the exploration cost comes as an important factor in assessing the quality of the search. Another problem is the need to approximate network semantic similarity measures. During the search process, only the partial network is revealed to the user, which renders faithful computation difficult. An important question is to find out the best way to leverage the currently available information to more accurately predict the real similar objects of interest.

With the abovementioned challenges in mind, the following subsection maps out the proposed method to cope with the research problem this study faces.

1.2 Methodology Overview

Given a heterogeneous network, a query entity, the node type of the resultant nodes, the objective meta-path semantics, and the metric that measures the meta-path semantics similarity, we propose to combine the notions of sampling a network by exploration[8][12] and retrieving a node set of similar objects (in the sampled network) based on the similarity metric dictated by the given meta-path semantics.

The major aim of this paper is to design a search framework by exploration that extracts the objects of the target node type given only gradually observed information through the sampling process. During the sampling process, the framework first evaluates the nearby neighbors, then selects suitable candidate nodes, and finally issues a request to access a particular candidate's information. In particular, we wish to estimate, based on the currently sampled subgraph, the expected similarity between the

query nodes and candidate nodes had we have the fully observed network at hand. We select the from the candidate nodes the node that matches our belief of semantic similarity the most.

Because type information for the unexplored portion of the network is not available, we propose to estimate the expected similarity using expected type distribution computed using the observed conditional probability between node types and link types [16]. Heuristics based on network based score propagation mechanism is proposed to incorporate type information to score approximation. We continue to sample until a desired number of objects of target node type is gathered.

1.3 Research Questions and Contributions

Given the scenario entailed in the proposed methodology, we summarize the research contribution of this paper below.

1. We first define the novel problem of semantic-driven searching entity in a heterogeneous social network that has only partially available network information, and the generic methodology that bases on sampling by exploration that solves the research problem.
2. Given the context of path-based semantics, as defined by the meta-path relation, we propose a method to incorporate currently observed meta-path statistics, and our belief of expected likelihood of node type distribution to determine how confident we should sample among the candidate nodes. In particular, we design a one-iteration mechanism to propagate scores to characterize meta-path connections between two nodes.
3. As for evaluation and experimentation, we test our methodology against several sampling baselines across a variety of real-world heterogeneous social

networks. We discover that our method is able to retrieve entities of interest using few nodes under a set of different meta-path semantics.

4. Lastly, we check the possible applications of using entity sampling by exploration. Scenarios tested on interpretability of our results and retrieval accuracy that connects to entity sampling by exploration are discussed. We further show model learning can be applied to align node instances with only attributes information and semantic similarity with the query.

1.4 Paper Organization

The rest of the thesis is organized as follows. Chapter 2 touches on relevant works to this work's research topic. In Chapter 3, we define the necessary preliminary background and formal problem definition used in the paper. Chapter 4 introduces our proposed framework of sampling and methodology of score estimation. Chapter 5 shows the experiment settings and implementation details. In Chapter 6, we discuss some observations and results that explain the nature of our method's superiority over the competing baselines. Chapter 7 presents some ideas on how the research concept in this work can be used to derive interpretable results. Chapter 8 is the conclusion and directions for future works.

Chapter 2

Related Works

For the problem of similarity search in heterogeneous social network via explorative sampling, two crucial components in the search process must be discussed: how one expands the answer set by exploring neighbors in the unexplored network, and how semantic-driven search of network nodes by traversing the graph topology is conducted.

2.1 Graph Sampling by Exploration

There are two main ways to represent the network: graph summarization and graph sampling. Graph summarization takes input the whole network and returns a condensed version of efficient visualization or processing of the graph[24]. However, as we do not have the access to the full network, that leaves us to graph sampling for solutions.

Representative subgraph sampling selects a subgraph via a choosing strategy. This method assumes that the network is only partially available and decisions must base on the currently sampled information. Usually, an approximation goal is presented, and the sampled graph is to be as close as the full observed network as possible in respect to that goal.

In the existing literature, some categories of sampling by exploration exist: for homogeneous networks, popular methods include sampling by Random Walk [1][4][9][17], Forest Fire[7], Ego-Centric Exploration[11], and Multiple Ego-Centric Exploration Sampling[8]. In these sampling methods, usually a set of seeds (which correspond to a set of node queries) is at first initialized to be the set of ‘egos’. At each

stage, the set of egos is maintained, and the sampling step looks at the immediate neighbors of the egos for candidate selection. Upon selecting the proceeding node from the set of candidate nodes, the ego in consideration is updated to the newly selected node. A major drawback for these sampling methods is that inter-type relationship and relative weighting for different types are not considered.

As for heterogeneous network sampling, [2] proposes a multi-graph approach that first decomposes the heterogeneous network into a set of homogeneous ones, each for a single type of link. Then for each single-type network, it performs random walk sampling before combining results. Again, the projection to homogeneous network falls into the same pitfall as homogeneous sampling methods. Two other relevant works include using Respondent Driven Sampling[8], and using Relational Profile Preserving Sampling[16]. The former approach, however, is not suitable for the task of search by exploration, as that research's aim is to estimate node type distribution using extracted subgraph. The latter approach, Relational Profile Preserving Sampling, starts out with a proposed network property, Relational Profile, which characterizes the conditional type probability distribution between any pair of node types derived from two connecting nodes. The selection process depends on how the candidate node fits the objective, or, minimizing Relational Profile difference between the sampled and original networks. This work bears some resemblance to our study scenario, as we need to assign weight according to the node's implied semantic. We differentiate ourselves in that the research's goal is not the same, leading to disparate scoring strategy.

2.2 Graph Semantics and Similarity Search

To more faithfully capture the characteristics of semantics denoted by a heterogeneous social network, structured representation of semantics is introduced.

Works that deals with information network usually play with node to node relations directly[6]. [20] discusses various aspects of graph schema, and how conditional type semantics can be estimated without the full knowledge of graph schema. Graph-denoted semantics, which allow type pattern to be structured as graph-like, are used in subgraph retrieval[19][21]. Often, these works require the access to a static database of large repository of graph patterns in order to mine frequent or similar graph portions that match the semantic pattern. One line of work that deals with a particular set of semantics is meta-path[14]. Meta-path confines the study of semantics to path connections of node types only. Studies reveal that meta-path can be used to derive many useful features for link prediction[22], clustering[23], similarity search[18], etc. Advantages for using meta-path are that the semantic increases interpretability, and simpler path structures are easier to manipulate, through optimization or random surfing.

As meta-path can measure the similarity between two objects based on how relevant they are connected according to the dictated semantic path, several measures have been tried to see the effectiveness of meta-path metrics on application performance. Simple path count, symmetric path count, normalized path count, and normalized symmetric count are among the early methods of measuring semantics similarity. Recently, a work[14] proposes a new metric, PathSim, that accounts the normalized meta-path counts with respect to the count that the end-nodes connecting back to itself. In addition, a fast solution to calculating PathSim is proposed using co-clustering to deal with large matrices.

The more general case of similarity search in heterogeneous social network are roughly divided into two subdomains. One methodology is to return a connected network, often using a variation of Steiner-Tree algorithm to connect query nodes. One

exemplary approach is MING[6]. In this framework, a connected subgraph is first discovered. Based on this connected subgraph, the algorithm computes the selection likelihood of the candidate node based on the probability that any random walker from the current subgraph can walk back to the subgraph via the candidate node. Score estimation build on the statistics on fact patterns, or the number of pairs of nodes that are connected by a particular relation (e.g. a membership relation like 'IsA'). This approach is quite different from our setting, as we are only interested in the set of objects in target type, not on how to connect nodes in the subgraph. Another methodology returns a set of similar objects of given node type. Unsupervised approach directly measures the similarity between two nodes using a given evaluating metric[14]. Supervised approach, on the other hand, requires the calculated similarity between two nodes of interest (source and target). After generating a list of instances and corresponding features, a ranking model is learned[18]. The supervised approach enjoys the advantage of feature engineering and objective function setting. However, both approaches do not apply in our scenario as we are restrained to acquiring information on the fly. A complete, offline database is not available beforehand.

In summary, there is no existing work that combines both paradigms of sampling by exploration and structural semantics in node similarity search in heterogeneous network. Our work succeeds in combining these two disparate aspects into solving the research problem.

Chapter 3

Problem Definition

3.1 Background and Terminology

Given a graph G = (V, E), where V is a set of N vertices (or entities, as we will interchangeably apply the term) and E represents a set of M edges (or relations), we define a **heterogeneous graph** as the following:

Definition 3.1: *a **heterogeneous graph** G = (V,E) is a graph, where each node, n, can be described by (n, NT(n)) that denotes: node n is of type or category NT(n), where a set of node type labels are denoted as NT. Each edge e can be described as a triple (v_1, v_2, ET(e)) that denotes: node v_1 and v_2 are related by relation ET(e), where edge type labels come from a set ET. Also, we use N(v) to denote neighbors of a node v.*

Referring back to the toy example in Figure 1-1, a publication network is clearly defined by several crucial entity types. In the network, *paper* publishes in *a journal; paper* cites other *papers; author* publishes *papers;* and *author* belongs to *organizations*. We can see that the semantic interpretation is almost deterministically described.

Also we take note of the notion of ***graph schema:*** a graph schema is a visualization of how a heterogeneous graph G is characterized by node type semantics NT and edge type semantics ET. In a graph schema, two node types nt_1, nt_2 are connected if and only if there exists edges in G that creates a path connecting two nodes (n_1, n_2) by types (nt_1, nt_2) respectively.

While edge type and node type can give intuitive and direct explanation to certain characteristics in a network, in order to describe higher-order semantics, we formally define structure semantics, meta-path, below:

Definition 3.2: *a **meta-path** $P = (T_1,\ldots,T_N)$ of N entity types (each T_i is a specific node type in NT) is a structural semantics where P denotes valid paths that relates two objects of types T_1 and T_N following the defined type connections. We call a string of nodes (entities) in G, $(n_1,\ldots,n_N)$ to be a **match** of meta-path P iff $NT(n_1) := T_1$, $NT(n_2):=T_2,\ldots,NT(n_N):=T_N$.*

This definition of meta-path ensures interpretability and reasonability whether the given meta-path semantic P pertains to graph G. If P is not a path allowable by the corresponding graph schema of G, then searching entities under P would be meaningless, or similarity based on P between two nodes of node type T_1 and T_N will always be 0.

We shall illustrate how to construe meta-path semantics through real-life example derived from the publication network in Figure 1-1. See Figure 3-1 for a set of meta-path patterns. For the path (APA), we are comparing similarities between two authors that share common paper. An equivalent view of this semantic is the strength of co-authorship between two authors. For the path (APCPA), we are interested in this case how two authors are connected through publishing papers at common venues. Take another path for example, (APC), we compare the similarity between an author and a journal venue through common papers. An equivalent interpretation would be how often an author publishes a paper at a given venue: the more frequent the author publishes, the more preferred the author publishes at a domain relevant to the journal

venue. Through these motivating examples, we find that any pair of nodes can be correlated by a certain meta-path, regardless whether or not the two nodes are of the same type, as long as the meta-path is supported both by the corresponding graph schema and human interpretable meanings.

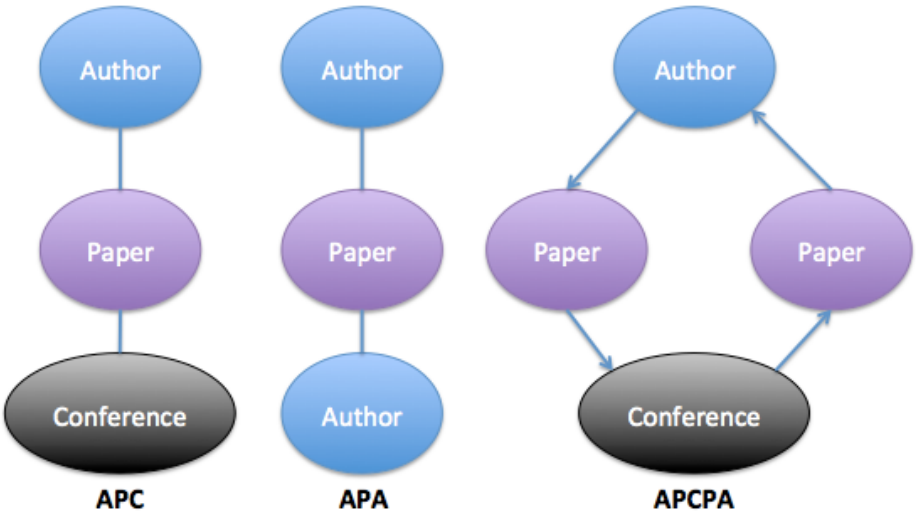


Figure 3-1. A sample of meta-path semantics

In order to measure the similarity between two nodes according to a particular meta-path P, we need a ***similarity metric*** $\mathfrak{M}(x, y)_P$ for actual computation. Some existing measures [14] include path count, which computes the number of path instances p between x and y following $\mathcal{P}$, or Random Walk, which measures the probability that a random walker starting from x ends with y following meta-path P. However, as suggested, in [14], symmetric similarity measure strengthens the belief between two nodes and provides better interpretability. We will explain more in detail below and use symmetric metrics as the main measures in our study:

- Symmetric random walk (SRW): this measure further entails that the probability requires random walker walking back from y to x, other than probability of walking from x to y, leading to the formulation of $\mathfrak{M}(x, y) =$

$\sum_{p \in P} Prob(P)$, where P = $(P_1 P_2)$: P_1 is path from *x* to *y* and P_2 likewise the path in opposite direction.

- PathSim: PathSim is a measure that further normalizes path count by pairwise random walks of the two nodes of interest. The exact computation supposes the function $\mathfrak{M}(x, y) = \frac{2*|\{p_{x \to y} : p_{x \to y} \in P\}|}{|\{p_{x \to x} : p_{x \to x} \in P\}| + |\{p_{y \to y} : p_{y \to y} \in P\}|}$. PathSim shows to have consistent performance of the metric over other metrics in interpretability[14]. Note that PathSim requires two end-nodes of the same node types to maintain certain properties proposed in the prior work.

We show by example how bi-directional metrics differ in semantic interpretability. Consider (APC) in Figure 3-1, $\mathfrak{M}_{SRW}$(x, y) extends the original explanation in that random walk probability in C-P-A paths denotes the author is a dominant author at the conference (which is different than the A-P-C interpretation).

Since PathSim is only defined over pairs of nodes with $T_1 = T_N$, so we use symmetric random walk to relate two nodes of different types in our experiments.

3.2 Formal Problem Definition

What we really want to achieve in our study is to return a small list of nodes that is the most similar set of nodes to the query node n_{seed}. That is, the sequence S of sampled objects under a particular meta-path semantic should have similar scores to the top-scored nodes in similarity using the full network. To compare the sampled node set S and V, all type T_N nodes in the full network, we use the likelihood function $\mathcal{L}(S, V)_P$.

Summarizing all aforementioned notions, we will provide the formal problem definition:

Given a particular meta-path P, seed node n_{seed}, size of entity set k, desired target node type T, and evaluating metric $\mathfrak{M}(x, y)_P$, the problem of ***meta-path semantic***

preserving entity search in heterogeneous social network is to sample sequentially from a heterogeneous network G a set S where |S| = k and NT(S) = T, with the relative scores between S and V (V = {n|nt(n) = T, n∈N}), $\mathcal{L}(S,V)_P$ maximized.

Chapter 4

Methodology

Our proposed algorithm bases on the idea of combining **explorative sampling** and **meta-path similarity score propagation.** In order to solve the formalized problem, we must first understand the nature of sampling by exploration to better describe the details of the proposed methodology.

4.1 Sampling by Exploration Framework

We briefly describe the process of sampling a set of entities in a heterogeneous network by exploration. Given a query seed node n_{seed}, the process is to sample a network G_s with a node subset S of target type T, where at each step during sampling, the new node n_{new} to be sampled is selected from the set of candidate nodes C_{Gs}, where C_{Gs} consists all unsampled one-step neighbors of the currently sampled network. As a new node is sampled with respect to G_s based on its immediate neighbors, we want to determine $argmax_S \mathcal{L}(S,V)_P$. However, to determine this set, we need to try out all possible combinations of node subsets, which is computationally intractable as the number of possible candidate sets grow exponentially large. To improve the efficiency, here an explorative sampling framework is defined as:

$$\forall v \in C_{Gs},\ P(n_{new} = v) \propto \mathcal{F}(v, G_s)_P = argmax_{S_{G_s'}} \mathcal{L}(S_{G_s'}, V)_P$$

where $G_s' = G_s + v$. The above equation essentially says the probability of a node to be selected as the next candidate to be sampled is proportional to the normalized score function $\mathcal{F}$ that represents how well including this node into the existing network can

match the desired semantics *P.* The process is shown in Algorithm 1.

Algorithm 1: Heterogeneous Network Entity Sampling by Exploration **Input:** seed node n_{seed}, k = sample size **Output:** Sampled entity set S $\subseteq$ V, graph $G_s = (V_s, E_s)$	
1	$V_s = \{n_{seed}\}$, $E_s = \{(n_{seed}, x) \mid x \in N(n_{seed})\}$
2	$S = \emptyset$
3	$n_{ego} = n_{seed}$
4	while $\|S\| < k$ do
5	$C = N(n_{ego})$
6	v = weightedSample($\mathcal{F}(v, Gs)$), $\forall v \in C$
7	$n_{ego} = v$
8	$V_s = \{V_s, v\}$
9	$E_s = E_s \cup$ edges(v)
10	End

This process addresses the problem of limited accessibility of the network. We are only granted the permission to retrieve nodes that are directly connected to the currently sampled subgraph. The sampling module will access the network only when the next node is selected out of possible candidate nodes.

4.2 Path Semantic-Aware Sampling for Heterogeneous Social Networks

Based on sampling by exploration, we propose path semantic-aware sampling (PASS) for heterogeneous social networks**,** where we try to ensure nodes of type of interest with highly similar semantics are sampled as early as possible. Recall that a big challenge to calculate the actual meta-path semantic similarity is that we do not know the type of a candidate node before it is sampled. In addition, from the user's

perspective, nodes that are not immediately connected to the sampled network are not accessible for processing. To resolve this challenge, we propose to predict the meta-path semantic similarity score distribution for each candidate node *v*, and use such distribution to generate the 'expected' meta-path similarity of *v* given the sampled network G_s. In particular, we plan to resolve the computation through two possible approaches: one by calculating the aggregated marginal value of meta-path similarity that measures the similarity between any given node and the query node; another by calculating all partial meta-path counts for all sampled nodes and propagate the partial path counts to candidate nodes of interest.

For partial meta-path count, the computation is relatively straightforward, as we are interested in counting $|n_{seed}\text{-}> x|_s$, $s \subset P$ for a specific meta-path P. As for computing $\mathfrak{M}(n_{seed},x)_P$ in general, where x is only an intermediate node in the meta-path, we give the following explanation. For all possible type(x) = T_i in position i of meta-path P (or partial meta-path path P'),

$$\mathfrak{M}(n_{seed}, x)_P = P((n_{seed} \to x | P', \mathfrak{M}))$$

Effectively, the calculated meta-path similarity metric for the intermediate node *x* is how the query node closely connects *x* based on the partial meta-path. This score is equivalently the marginal probability that n_{seed} travels to *x* dictated by the partial meta-path P' and measured in respect to $\mathfrak{M}$ Simply put, the score $\mathfrak{M}(n_{seed},x)_P$ computes how *x* contributes on average to the computation of $\mathfrak{M}(n_{seed},v)_P$, where type(n_{seed}) = T_1 and type(*v*) = T_N, at the either ends of the meta-path semantic, for any node *v* satisfying the node type specification.

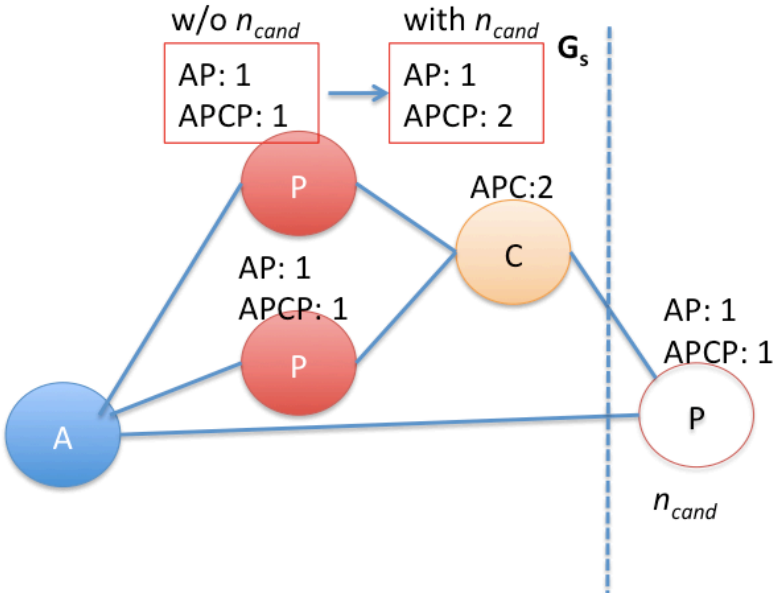


Figure 4-1. An illustration to how to compute $\mathfrak{M}(v, n_{seed})_P$ in G_s

Readers may refer to Figure 4-1 for example. In the illustrated example, we use the query meta-path (APCPA). Consider candidate node v_1, with the node type = Author. In this case, we simply compute $\mathfrak{M}(v_1, n_{seed})_P$ according to suitable type-type adjacency matrices computation. For v_2, of type = Paper, suppose the node is positioned at T_4 of the meta-path, we then have all $\mathfrak{M}(v_2, n_{seed})_{(APCP)}$ computed; we consider scores for all connecting nodes for v_2 and $n_{type=Author}$ as well.

Having defined property methodology for computation for the semantic metric, we now turn to explaining likelihood function in terms of mathematical formulation. We wish to sample based on:

$$\mathcal{L}(S_{G_s+v}, V)_P \approx \mathfrak{M}(n_{seed}, v)_P + \mathcal{L}(S_{G_s}, V)_P$$

$$\Rightarrow argmax_{S_{G_s'}} \mathcal{L}(S_{G_s'}, V)_P = argmax_{S_{G_s'}} \mathfrak{M}(n_{seed}, v)_P + \mathcal{L}(S_{G_s}, V)_P$$

$$\Rightarrow argmax_{S_{G_s'}} \mathcal{L}(S_{G_s'}, V)_P = G_s + argmax_v \mathfrak{M}(n_{seed}, v)_P$$

$$\therefore \mathcal{F}(v, G_s)_P = \mathrm{E}[\mathfrak{M}(n_{seed}, v)_P | G_s, \mathfrak{M}(n_{seed}, v)_P | n \in N_s\}]$$

where G_s' is $G_s + v$, and N_s is the node set for the sampled network G_s. Because the information from the full network is always invisible for access, we adopt incremental

estimation for likelihood. We can also view the estimation as the probability of *v's* contribution to the calculation of $\mathcal{L}(S_{G_s}, V)_{\mathrm{P}}$. Since the likelihood is estimated by node-wise approximation of meta-path similarity metric, it effectively becomes the node selection score distribution. Note that we have relaxed the deterministic selection to probabilistic sampling to avoid overfitting local structure.

Having defined what to compute for $\mathcal{F}(v, G_s)_P$ and how to compute $\mathfrak{M}$, we now wrap up with how to deal with estimating information from the unsampled portion of the network. We propose to compute the metric score into two stages: one to estimate $\mathfrak{M}(n_{seed}, v)_{\boldsymbol{P_{1:i}}}$ while estimating the expected node type distribution for node *v*; one to update by propagating the marginal score down the network. The following section individually solves the problem of how to calculate the probability, with an approximation methodology.

In fact, if we further separate node *x* from the multiplication, we have:

$$\mathfrak{M}(n_{seed}, x)_{\boldsymbol{P_{1:i}}} = \sum_{n \in N(x)} \mathrm{I}[\big(\mathrm{type(n)}, \mathrm{type}(x)\big) = (T_{i-1}, T_i)] * \frac{\mathfrak{M}(n_{seed}, n)_{\boldsymbol{P_{1:i-1}}}}{\mathrm{Z}}$$

where Z is the normalization factor depending on the metric used and the sampling method applied. Note that the indicator function dictates the type transition to match meta-path semantic. It is straightforward to see that the above combination of scores from neighboring nodes should produce $\mathfrak{M}(n_{seed}, x)_{\boldsymbol{P_{1:i}}}$, as expected.

4.3 Score Estimation Strategy

The entire sampling heuristic requires us to know how

$$\sum_{n \in N(x)} \mathrm{I}[\big(\mathrm{type(n)}, \mathrm{type}(x)\big) = (T_{i-1}, T_i)] * \frac{\mathfrak{M}(n_{seed}, n)_{\boldsymbol{P_{1:i-1}}}}{\mathrm{Z}}$$

is computed given the node of interest *n* and the sampled network. We break this into two parts: expected type probability estimation and metric score estimation.

4.3.1 Expected Node Type Estimation

In short, we are interested in $P(type(v) = t|G_s)$, which is the expected node type distribution given currently available type and topology information.

To conduct the calculation for type estimation, node type of *v* can be estimated from its neighboring nodes. This is because naturally certain types of nodes are connected to certain types of node/links. Mathematically, we denote the node type contribution as dictated by the connected neighbors of the node *v*: $P(type(v) = t|G_s)$ = $P(type(v)=t|type(N_v))$, where type(N_v) represents the type information of observed neighbor edges and nodes of *v*. That is, N_v is the joint set of observed links and nodes connecting to *v*. Using Bayes rule, we can transform $P(type(v)=t|type(N_v))$ into $P(type(N_v)|\ type(v)=t)$ *$P(type(v)=t)$, since the $P(type(v)=t)$ is deterministic (i.e. probability 1) as they are already observed Using Naïve Bayes assumption to assume the type of each neighboring node/link given type(*v*) is independent, we can further decompose $P(type(N_v)|\ type(v)=t)$ into the following:

$$\prod_{i \in N_v} P(type(i)|type(v) = t)$$

$P(type(i)|type(v) = t)$ can be obtained from the empirical conditional type counts of G_s, since the type of observed node i is already known.

4.3.2 Metric Score Estimation

Based on the intuition presented earlier, we entail our first heuristic to represent meta-path similarity metric estimation, which is an **one-iteration propagation,** implemented in multiple **depth-first-search runs**, name as **MetaProp** similarity scores for the sampled network G_s, where we initialize n_{seed} to the value 1 and propagate the

score down G_s for one iteration through the following formulation. Denoting the propagation scores for node v is $M(v)$, we have:

$$M(v) = \mathrm{E}_{type(v)}[\sum_{n \in N(v)} S(n,v) * \frac{M(n)}{\deg(n)}]$$

$$= \sum_{t} P(type(v) = t | G_s) \sum_{n \in N(v)} S(n,v) * \frac{M(n)}{\deg(n)}$$

$$= P(type(v) = t, t = T_{i+1} | G_s) \sum_{n \in N(v)} S(n,v) * \frac{M(n)}{\deg(n)}$$

where T_{i+1} represents the type at i+1th position following the node type and position in the meta-path occupied by n. The calculation of the score is due to the propagation of scores from the already observed neighbors of v, over all possible node types that v may hold. $S(n,v) * \frac{1}{\deg(n)}$ is the transition function from type(n) to type(v). For the case of this study, we leave out the weighting function $S(x,y)$ and let the transition behavior completely determined by the neighboring degree. Since $\sum_t P(type(v) = t | G_s)$ has been explained earlier, so we focus on the latter portion, $\sum_{n \in N(v)} S(n,v) * \frac{M(n)}{\deg(n)}$. We take important note that this propagation bases heavily on random walk implementation to ensure the score is not fixed to some local topology, thus overfitting the locality and mis-estimate the global structure. A complete score estimation can be seen as:

Combining the terms together, $\mathcal{F}(v, G_s)_P$ can be approximated as:

$$\prod_{i \in N_v} P(type(i) | type(v) = t) * P(type(v) = t) \sum_{n \in N(v), n \in G_S} s(n,v) \frac{\mathfrak{M}(n, n_{seed})_P}{\deg(n)}$$

$$\Rightarrow \mathcal{F}(v, G_s)_P \cong \prod_{i \in N_v} P(type(i) | type(v) = t) P(type(v) = t) * \sum_{n \in N(v), n \in G_S} s(n,v) \frac{M(n)}{\deg(n)}$$

We venture to explain further. Note that since we are propagating scores down the network by multiple runs of rooted random walk, it is likely to travel a node many times and updating the value based on the neighborhood. In effect, this is to aggregate the current marginal statistics of the neighbors, which adds influence of computation by other nodes of type T_N. This formulation clearly benefits meta-path metrics that favors symmetric relations, such as Symmetric Random Walk or Pathsim. We prefer this behavior because optimizing bi-directional metrics clearly favors generating interpretable results, as demonstrated by the experiments in [14].

The merit of this formulation is that in the case of large-scale network, this formulation easily blends in with existing state-of-the-art computing methods for matrix computation. Also, the proposed method avoids the problem of needing to constantly maintain multiple partial semantic similarity matrices.

We define the generalized case for the propagation measure that spans multiple iterations, which can be calculated as:

$$M(v) = \frac{1-\alpha}{|V_S|} + \alpha \sum_{n \in N(v)} S(n, v) * M(n)$$

Here $S(n, v)$ is slightly modified to include division by node degree (which all effectively fuses to transition probability). Now we wish to show that this formulation leads to convergence, thus proving the usability in any case of problem setting.

Theorem 4-1. Given a sampled network G_s, fixed matrix S(a, b), and $\alpha = [0, 1)$, *following expansion calculation of scores in initialization,*

$$M(v) = \frac{1-\alpha}{|V_S|} + \alpha \sum_{n \in N(v)} S(n, v) * M(n) \text{ converges.}$$

Proof. We first complete the score initialization phase. Then,

$$\forall n \in N \text{ in } G_s, M_0(n) = n_0$$

, as some form of calculated score through propagation. So, for every node , there is a corresponding score n_0. We put all scores into a score vector $V = [v_{10}, v_{10}, \ldots, v_{|N|0}]$. Now we have:

$$M(v) = \frac{1-\alpha}{|V_S|}V(v) + \alpha \sum_{n \in N(v)} S(n,v) * M(n).$$

Since all scores in *V, S, and* $\mathfrak{M}$ *are fixed, and between the interval of [0,1], Following the reasoning in [6], we see the scores will be positively recurrent and aperiodic, satisfying ergocity.* ■

We take note that this generalized means to summarize the propagation effect, under the constraint of meta-path search length and type, is similar to the Propflow predictor[30]. However, we differ from Propflow in that they are interested in *l*-step propagation in a Breadth-First-Search manner, and do not apply random surfer model as we need process newly included information one by one. In addition, no constraint is imposed on the case of meta-path similarity search for Propflow.

There is, however, some limitations to the DFS approach. This is because the scores represented for each node is effectively some aggregated score for different partial meta-paths. This may not convey the particular meta-path score that we are really interested in. Instead, we further show a **Breadth-First-Search** approach, generalizing as **PAth Semantic aware Score propagation (PASS)**, explained as follows.

See Figure 4-2, the calculation is divided into three components: a recorded path count for each node in the sampled network; expected type probability for the node in interest, and path completion probability that matches the partial meta-path to the actual meta-path dictated in search criteria.

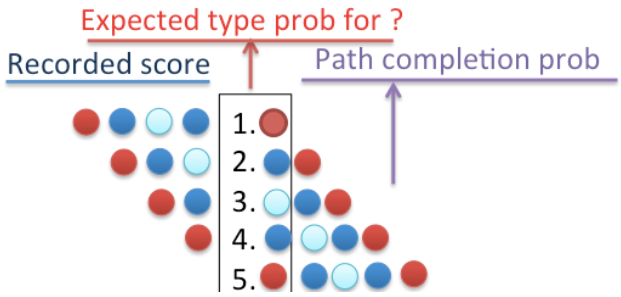


Figure 4-2. An illustration on the overarching concept to how to compute $PASS$

when used in a real case, as exhibited in Figure 4-3, we first compute the expected node type distribution for a particular candidate node given the sampled network. Using the computed type distribution and the existent type conditional probability table given by the sampled network, for any particular position and node type in the meta-path the candidate node may be, we compute the probability that the candidate node follows a particular partial meta-path. For example, in the publication network and given a meta-path (Author-Paper-Conference-Paper-Author), if we are interested in the partial meta-path that completes the query meta-path (Conference-Paper-Author), we compute the probability that the candidate node is of "Conference" node type, and use the conditional probability table to calculate the combined transition probability to Paper and Author nodes. Lastly, since we have a count table of all partial meta-paths (e.g. Author-Paper) for each node we sample, we can propagate these path counts to the candidate node, summing up all possible partial meta-paths (each multiplied by the completion probability). This will leave us the combined selection function, with the newly introduced terms labeled in red:

$$\mathcal{F}(v, G_s)_P \cong \sum_{n \in N(v)} \sum_{i \in pos(P)} C(n,i) * \prod_{i \in N_v} P(type(i)|type(v) = t) P(type(v) = t)$$

$$* \prod_{s=i+1}^{len(P)-1} P(type(P_s)|type(P_{s+1}))$$

where C(n,i) is the corresponding count table for node n and partial meta-path from position 1 to position i. This method provides us a better intuition as to how to directly combine different partial paths, while penalizing for greater uncertainty, for a possibly fairer weighting of all contribution to meta-path computation. This avoids the ambiguity in the previous DFS-based MetaProp algorithm.

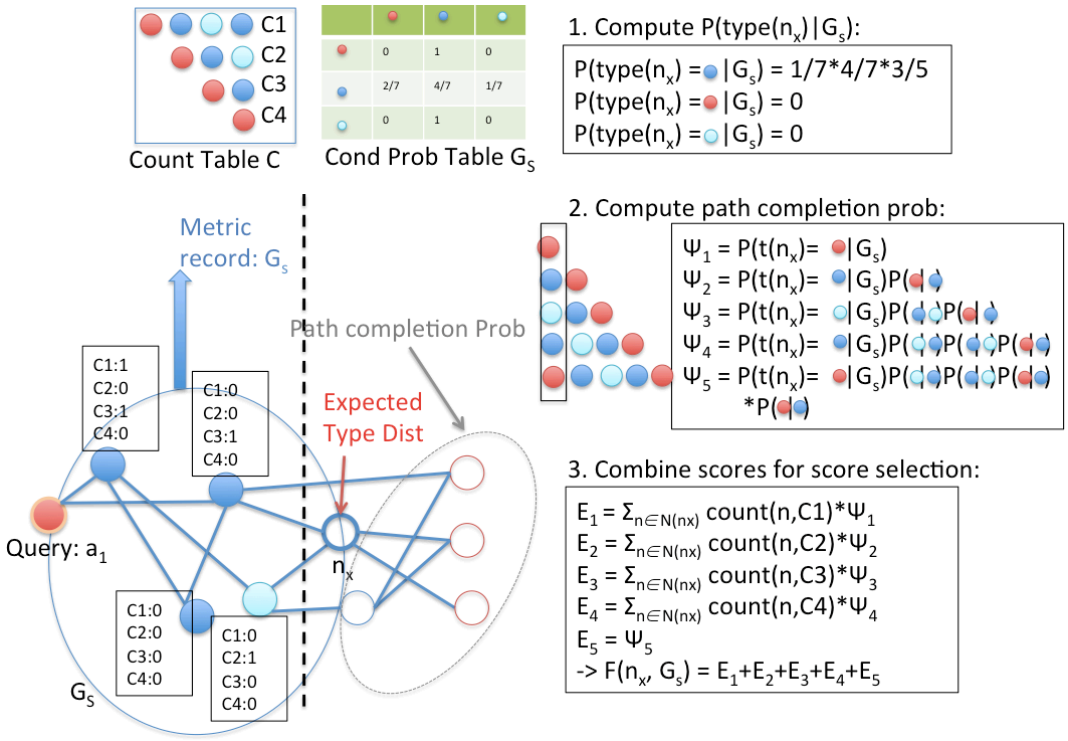


Figure 4-3. An example applying PASS, using publication network

To sum up, existing sampling by exploration methods (by random or topology-driven selection) suffer from problems of easily overfitting local graph structures. However, in face of heterogeneous graph, where certain types may be predominant in the graph, naïve sampling may miss some nodes that are sparsely connected, but contain strong semantic meanings. Instead, the proposed score sampling heuritics prioritizes meta-path semantic similarity approximation, ensuring each inclusion of a new node can recruit the most semantically similar node in the limited sampling space.

Chapter 5

Datasets, Experiments, and Evaluations

In this section, we will first explain the nature of the three real-life heterogeneous social networks used for the experiments of this study. In addition, we will discuss means for evaluating the performance of entity set search.

5.1 Datasets

Each of the three real-life heterogeneous social networks covers different types of social relations. The three networks are as the following: High Energy Physics Citation Network, DBLP Publication Network, and Move Network. The basic statistics of the networks is shown in Table 5-1, and the details for the origins and characteristics of the networks are shown in the ensuing subsections. We use the largest connected component from the extracted network data for the experiments.

	# Node	# Edge	# Node Type	# Edge Type
High Energy Physics Citation	41,744	483,217	4	5
DBLP Network	86,535	81,255	3	3
Movie Network	24,362	44,396	5	6

Table 5-1. Statistics of datasets.

5.1.1 High Energy Physics Citation Network

High Energy Physics Citation network[3] contains papers published from 1993 to 2003, which was released in 2003 KDD Cup. The network covers four node types: Physics Papers, Authors, Journals the Papers are published in, the E-mail Domain for the author membership. The edge types denote the citation relationships: Authored (i.e. Author *Authored* Paper), Published in (i.e. Paper *Published in* Journal), Cites (i.e. Paper *Cites* Paper), Email affiliated (i.e. Author is *Email affiliated to* E-mail domain). The corresponding graph schema for the network can be seen in Figure 5-1.

The nature of the publication network, we take note that there is a high imbalance in node type labels distribution for nodes, which may make pure walking methods prone to bias to certain node type, such as journal type.

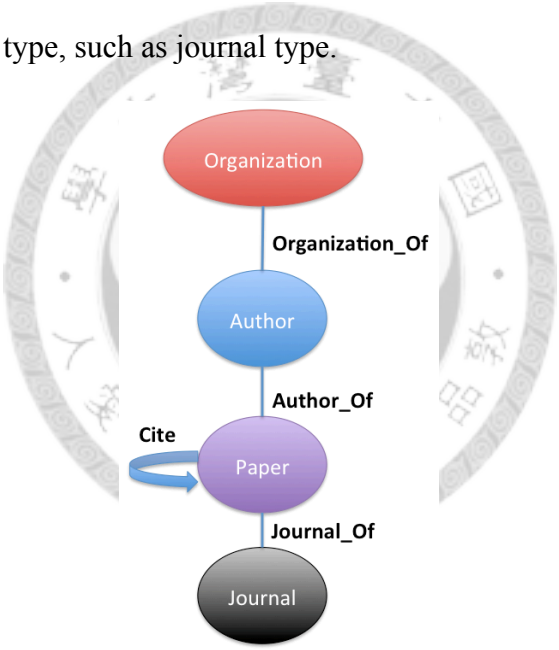


Figure 5-1. The network structure for High Energy Physics Citation Network.

5.1.2 DBLP Publication Network

For the DBLP Publication Network[4], it is a network of online repository of publication information for the research domain of computer science, made available by Trier University. We extract information from Artnet (Arnetminer.org), which provides

[3] http://www.cs.cornell.edu/projects/kddcup/

[4] http://www.informatik.uni-trier.de/ ley/db/

more structured data to explain the network. We follow the definition suggested in [14] to extract a subnetwork from the original data, using conferences ranked top in Microsoft Academic Network[6] for four major research areas: *Database, Machine Learning, Information Retrieval, and Data Mining*. We select a total of 22 conferences in our study (see Table 5-2 for the complete list). The network shares very similar semantic structure as High Energy Physics Network, as the node types are Paper, Author, and Conference/Journal. The link type dictates several relations: Authored (i.e. Author *Authored* Paper), Published in (i.e. Paper *Published in* Journal), Cites (i.e. Paper *Cites* Paper), Appears in (i.e. Word *Appears in* Paper abstract/title), Co-occur (i.e. Word *co-occur* with Word in certain number of papers). The complete graph schema of the network is shown in Figure 5-2. Again, take note of the sparse connections of intra-type links and high imbalance of the node labels.

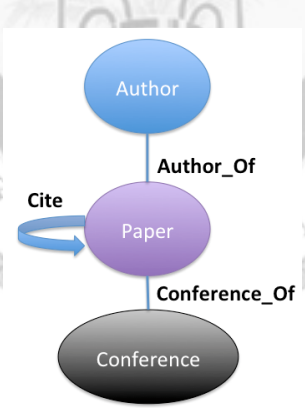


Figure 5-2. The network structure for DBLP Publication Network.

Database	Data Mining	IR	Machine Learning
VLDB	CIKM	SIGIR	ICML
SIGMOD	SIGKDD	WWW	UAI
PODS	ICDM	TREC	NIPS
ICDE	PKDD	ECIR	ECML
ICDT	PAKDD		IJCNN
EDBT	SDM		COLT

Table 5-2. The four research domains and corresponding conferences for the DBLP Publication Network.

5.1.3 Movie Network

The movie network is extracted from the imdb online movie database[5]. A rather complex connotation of entities and relationships exist between movie and contributors. For the sake of simplicity, we consider the type representations and interactions as dictated in Figure 5-3. For node types, we consider Actors/Actress, Movie, Contributors (Directors, Producers, and Writers), Place (Company of production), Year, Category (type or genre of the film), and Duration (the categorized length of movie in theaters). For link types, the network is characterized by several relations: Act in (i.e. Actor/Actress *Act in* Movie), Contribute (i.e. Contributor *Contribute* to a Movie), Made In (i.e. Movie *Made in* Place), Produce In (i.e. Movie *Produce In* Year), Belong To (i.e. Movie *Belong to* Category), and Last (i.e. Movie *Last* Duration). For this network, we observe a more balanced distribution for different types of people (namely, actors and contributors). Other attributes (duration, year, place, category) are more imbalanced.

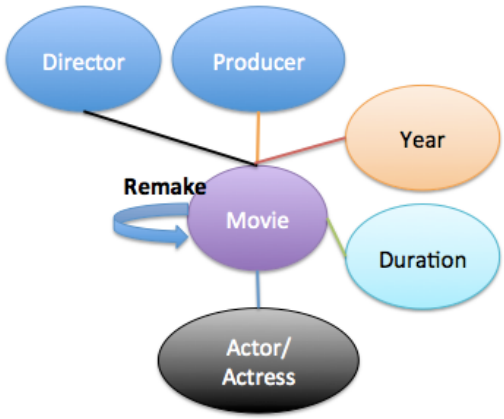


Figure 5-3. The network structure of Movie Network.

5.2 Evaluation Metric

In order to measure the relevance of the retrieved list and full network, we apply

[5] http://www.imdb.com/

the notion of comparing ranks of computed semantic similarity scores between entities of the two lists. The idea follows in that objects with high semantic similarity in the original network should be returned by the search algorithm while the less similar objects should not be returned by the algorithm until later stages of sampling. Therefore, to compare the similarity of meta-path semantics, we utilize the concept of **Normalized Discounted Cumulative Gain** (NDCG) as our evaluating measure. To calculate NDCG@k that concerns top-k most similar entities of type T to query Q, we impose a small modification. Before we apply NDCG directly, we sort the two sequences (S for the sampled nodes and $V_{type=T}$ denoting all nodes in G of type T) in terms of the metric $\mathfrak{M}$ to compare the strengths between the most similar nodes in each node list with respect to Q. To summarize, we arrive at the following computation for NDCG@k:

$$NDCG(S, V_{type=T})@k = sorted(\sum_{i=1}^{k} \frac{2^{\mathfrak{M}(si,Q)}-1}{log_2(i+1)}) / sorted(\sum_{i=1}^{k} \frac{2^{\mathfrak{M}(Vi,Q)}-1}{log_2(i+1)})$$

5.3 Baselines

For baselines in comparison, we consider several popular sampling methods:

1. **Random Surfer/Random Walk Model with restart:** equivalently viewed as PageRank measure[25], the sampling method tries to have a random walker keep moving to a neighboring node and with a probability restarting at n_{seed} with damping factor α. With initialized node score function δ, the importance scores can be evaluated as:

$$PR(v) = (1-\alpha)\delta(v) + \alpha \sum_{n \in N(v)} PR(n)$$

2. **Maximal Entropy Random Walk:** this sampling model emphasizes on creating equal probabilities for different paths connecting any pair of nodes in a network[26]. By enforcing equal path probabilities, we assume the imbalance of

distribution of nodes/edges across different labels may be rightfully reweighted. The effectiveness of such notion in relating node similarity has been applied to link prediction[27]. The explicit calculation for the transition matrix from i to j, P_{ij}, is as the following:

$$P_{ij} = \frac{a_{ij} * u_j}{\lambda u_i}$$

with adjacency matrix $A = \{a_{ij}\}, \lambda$ is the largest eigenvalue of A, and **u** is the corresponding eigenvector. As we do not have the full graph adjacency matrix beforehand, we use degree-based estimator for approximation[26].

3. **Locality Connected Random Walk:** this method focuses on maximizing connection back to sampled network. The more densely connected the candidate node to the sampled network, the more likely it is going to be selected [28]. We define the scoring function as:

$$\mathcal{F}(n) = \frac{\#\ edges\ (v,n), v \in V_s}{\#\ candidate\ nodes}$$

5.4 Implementation

For implementation of all sampling methods, we tried sampling from all one-step neighbors to G_s or just from the one-step neighbors of one n_{ego}. Experiment results show similar results for both approaches. To mitigate the computation cost for dealing with a large amount of one-step neighbors to G_s, we adopt random surfer model for all sampling models, introducing damping factor α that indicates restart probability at the starting query node n_{seed}.

For the implementation of the experiments, we chose to use networkx[6] library in Python language. In addition, MATLAB is used for some matrix computation. For

[6] http://networkx.github.io/

predictive model training, we use for learning ranking model. All experiments are run on a Linux server of cluster machines with AMD Opteron 2350 2.0GHz Quad-core CPU and 32 GB memory.

Chapter 6

Experiment

We will, in this section, examine different aspects of our problem setting and solution on the three different real life heterogeneous social networks described earlier. We first develop our intuitions for the proposed framework through some motivating experiments. Later, we will discuss results from various other settings, showing the superiority of our method. For the settings of parameters, we use restart probability = 0.15 for RWR (as suggested by [7]), DegRW, and LocalRW. All experiments are repeated on random seeds for five times to ensure stability in results.

6.1 Effects of Information Inaccessibility, Placement of Interest, and Meta-Path Metrics

We first begin our experimentation to see how preliminary settings of the research problem and evaluation method may affect the result.

Before showing experiment results, a list of meta-path semantics is listed in Table 6-1. We label the meta-path semantics with the corresponding node type labels dictated by the network (in the case, High Energy Physics) and briefly discuss the semantic interpretations. For the experiments, we concern with two nontrivial meta-path semantics that map to DBLP and Movie Networks: APCPA (i.e. Author- Paper - Conference/Journal-Paper-Author), which explains how two authors are related by the common publishing venues, and CPAPC (i.e. Conference/Journal – Paper –Author -Paper-Conference/Journal), which measures the similarity between publishing venues

by how many authors tend to publish at both venues.

Meta-Path Semantics	Explanations
Author-Paper-Conference / Journal-Paper (APCP)	Papers published at Author's publishing venues
Author-Paper-Author (APA)	Co-authorship
Author-Paper-Journal-Paper-Author (APJPA) ((APCPA) in DBLP and (AM[D/W/P]MA) in Movie)	Authors that share Publishing Venues
(APCPAP)	Papers by Authors that share Publishing Venues
$(APCPA)^2$	Authors that share Authors that share Publishing Venues
Conference-Paper-Author-Paper (CPAP)	Papers by Conference's publishing Authors
Conference-Paper-Author-Paper-Conference (CPAPC)	Conferences sharing Publishing Authors

Table 6-1. A list of meta-path semantics used in High Energy Physics Publication Network to compare similarity search via exploration of different pairs of node types. For each meta-path, the table provides a physical meaning of the semantic path. The highlighted meta-paths (APCPA, CPAPC) are used as common meta-path semantics throughout all three testing real-life networks, as labeled in red and orange.

We first ask ourselves a simple question: do our proposed methods actually work in situations where node types are visible? This is effectively a weakened constraint on our original problem definition. We should expect our methods' performance achieve competitive results in terms of similar node entity retrieval. Later, we will emphasize the idea of the significance in the difficulty of using 'guesses' on portions of network

where no apparent information is available. We plot the results for sampling with neighboring nodes' type information visible. For the baseline sampling methods, we enforce restart if the currently traversed node by the random surfer is not matching with the schema. See Figure 6-2 for the result in High Energy Physics Network using APCPA meta-path.

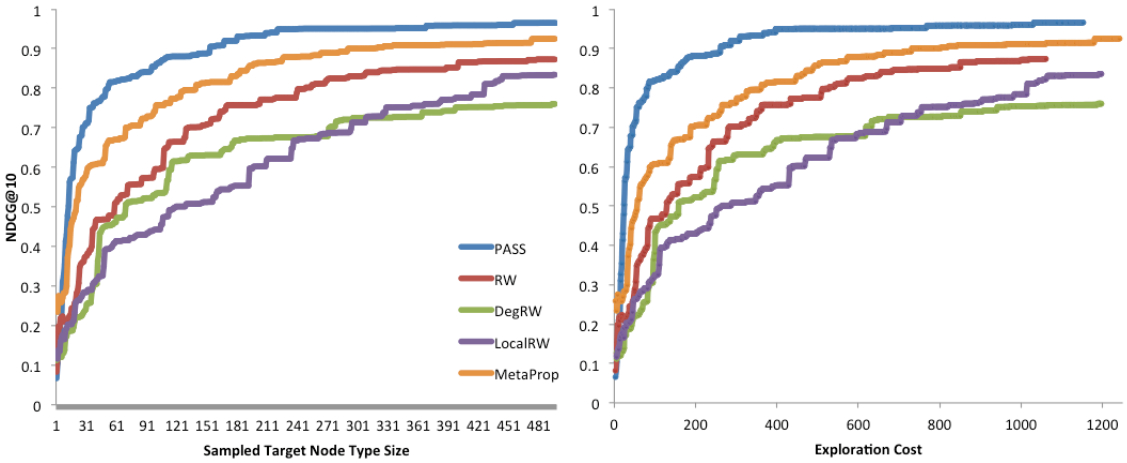


Figure 6-2. Results of Search Methods Under Visible Neighboring Type Information. Left: # of nodes sampled for target type v.s. NDCG@10. Right: # of total nodes sampled v.s. NDCG@10

We show the results of |S| v.s. NDCG@10 in the left and exploration cost v.s. NDCG@10 on the right, where exploration cost is the total nodes in G_s. The left figure is primarily for the purpose of showing how fast can the sampling methods achieve high NDCG given that users are sensitive at the cost of retrieving target type nodes. For the right figure, in each of the sampling methods, we sample nodes of the target node type for 500 nodes, or until the curve approaches an asymptotic bound. In this case, we can see that the proposed MetaProp and PASS spend about the same cost (a little bit more for MetaProp) exploring. However, we see that our proposed method reaches high NDCG accuracy at the early stages of sampling, because of the semantic similarity requirement. For other sampling methods, we see Random Walk with Restart is the strongest baseline. We can observe that as the pure Random Walk with Restart is not

biased, surfing nodes to different types will be easier, leading to lower cost. Topology biased methods tend to look to less relevant nodes, like those that are more closely connected to popular entities, but do not match the required semantic meaning. In short, we have shown that the simpler problem of visible node type exploration can be solved using our method in a very straight-forward way.

Next, we look at the results of inaccessible neighboring node type information. The results are shown in Figure 6-3 for High Energy Physics Network with APCPA meta-path semantic. For the search accuracy, without access to node type information makes sampling harder to estimate. The evidences can be found in the lowering accuracy in the sampling methods. However, even under this constraint, the proposed method still outperforms all baselines. Note that Random Walk with Restart is still the most competitive method, while other biased sampling by exploration methods can easily walk to nodes of high semantic dissimilarity to the query node and stuck to that local topology. MetaProp and PASS enjoy another advantage: since the sampling method can predict candidate node's label, a lot of exploration cost can be saved, as long as we can correctly infer the node type label. In both cases, PASS clearly performs superior in terms of NDCG measure evaluation.

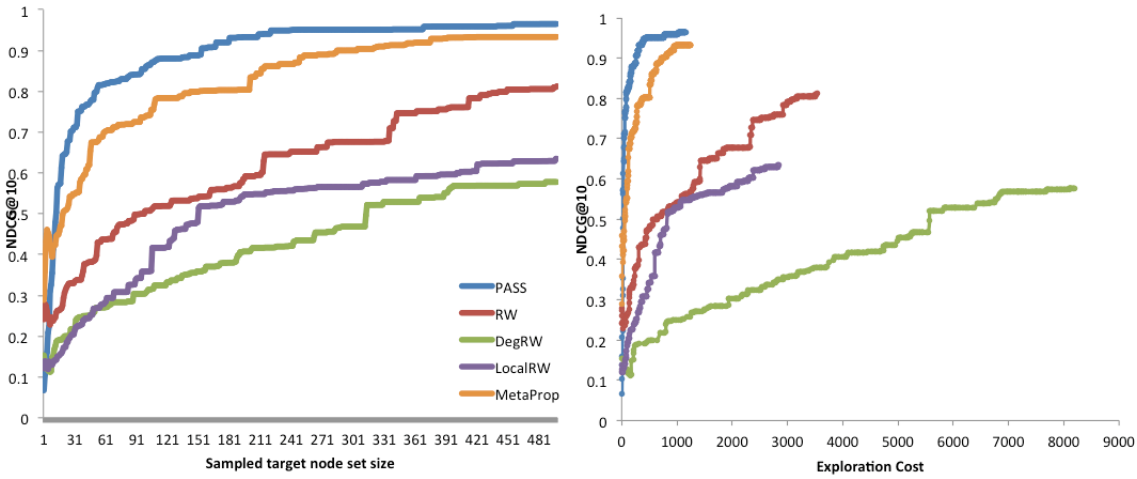
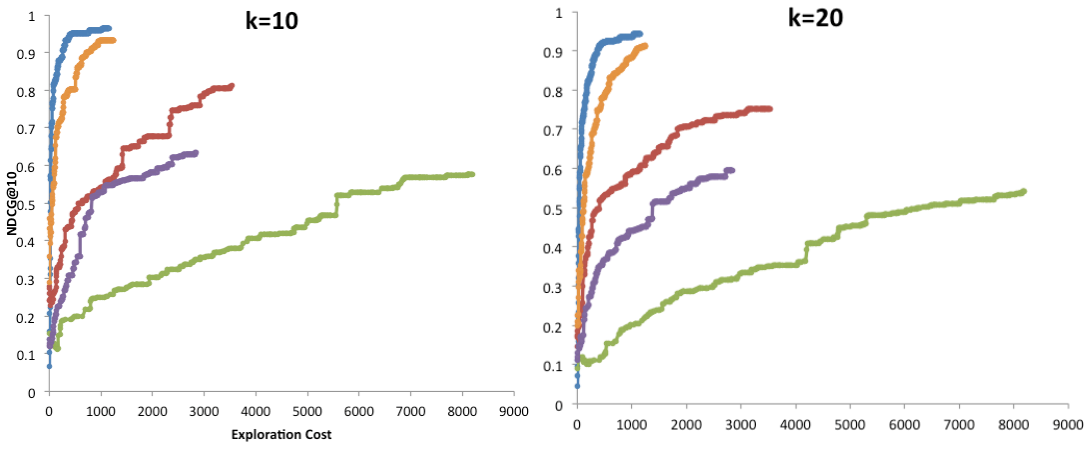


Figure 6-3. Results of Search Methods Under Invisible Neighboring Type Information. Left: # of nodes sampled for target type v.s. NDCG@10. Right: # of total nodes

sampled v.s. NDCG@10

Now, we look further into the nature of the evaluating metric used in the experiment: Normalized Discounted Cumulative Gain for sorted meta-path similarity scores. We are interested in the effect of k, or the number of top-ranked items of interest, that affect the measure of NDCG@k. The results of High Energy Physics Network are plotted in Figure 6-4. The experiments investigate four different kinds of k: k=10, k=20, and k=50. Such k are selected to reflect how many items a person would normally concern to determine similarity for a Author, Paper, Conference, Movie, etc. We observe that as k increases, the NDCG@k scores decrease in return. For the baseline methods, the increments decrease much more swiftly than the proposed method. For MetaProp, the performance also drops with increasing k. The only exception is the proposed PASS algorithm. The performance retains high NDCG for different k values. This shows that important nodes that hold high semantic similarity to the query node are sampled early in the proposed heuristic. Based on this concept, we can search for a safe threshold to sample, which is around |S|~150 for reasonable k~[10,50]. This conclusion fits intuition, as user usually is interested in searching only dozens of potential similar entities and too high a k may affect interpretability.



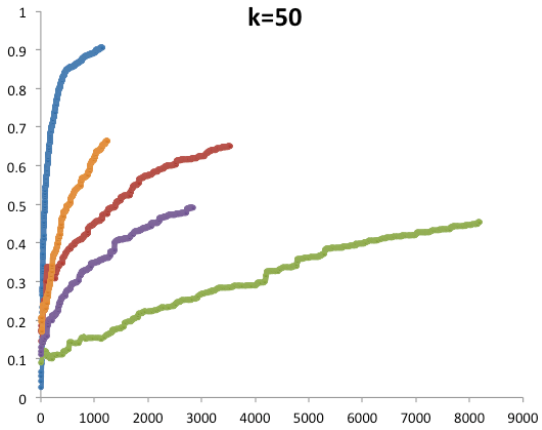


Figure 6-5. Plotting sampled nodes for desired object set S against NDCG@k for different values of k, using the High Energy Physics Network as example.

Lastly, we attempt to confirm the sampling NDCG results against different evaluating meta-path metrics are in accordance with the initial aim in the methodology design. See Figure 6-6 for results in both Symmetric Random Walk (SRW) and Pathsim metrics (sampling until |S|=500). We can see that different sampling methods perform differently when evaluating using different metrics. Most notably, we see Random Walk perform much worse in SRW metric, which is evident as methods like Local RW concerns with most expansion connection, which may introduce more connecting meta-paths back to n_{seed}. We note though, however the variability in performance, MetaProp consistently outperform both in terms of NDCG and in exploration cost.

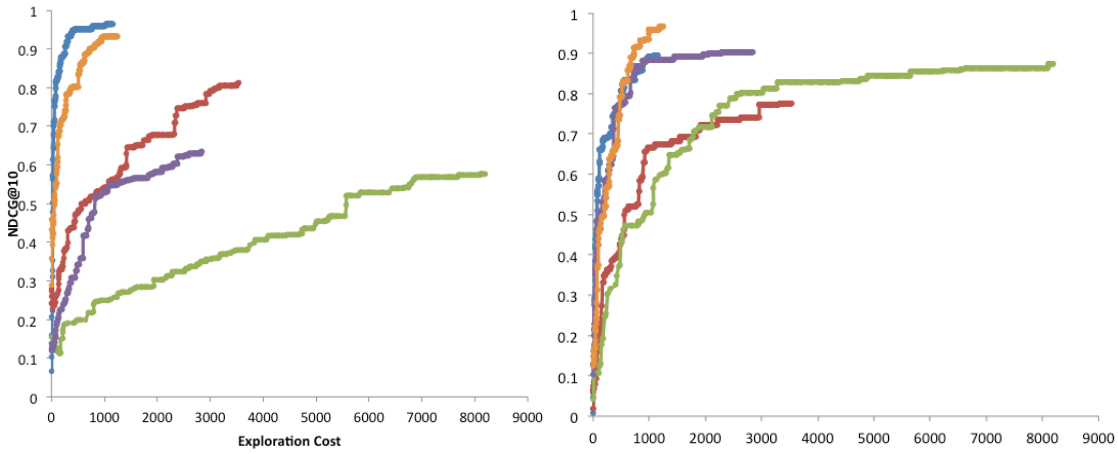


Figure 6-6. Exploration cost v.s. NDCG@10 for Meta-Path = (APCPA) in High Energy Physics Network. Left: Pathsim metric, Right: SRW metric. Sample till |S| = 500.

It is interesting to note, though, that PASS holds performance not much superior to the baseline methods, and starts to perform worse with more exploration cost than MetaProp. This reveals an important argument: there is a tradeoff between Breadth-First-Search sampling paradigm and Depth-First-Search sampling paradigm, depending on application. In DFS approach, multiple reruns allow random walker to continuously renew its belief on the aggregative count of partial meta-paths. This combination of multiple paths alleviates the problem of one-way path metric approximation. In the BFS approach, however, the objective focuses solely on one-directional path count of meta-path, which benefits Pathsim (as the numerator of Pathsim is the most important component), but not for heavily normalized measures like Symmetric Random Walk.

Based on the above experiments, for rest of the study, we set the parameters to: k = 10, and measuring metric: PathSim for node pairs of same type and SRW otherwise. The tradeoffs between BFS/DFS methods are displayed through different meta-paths.

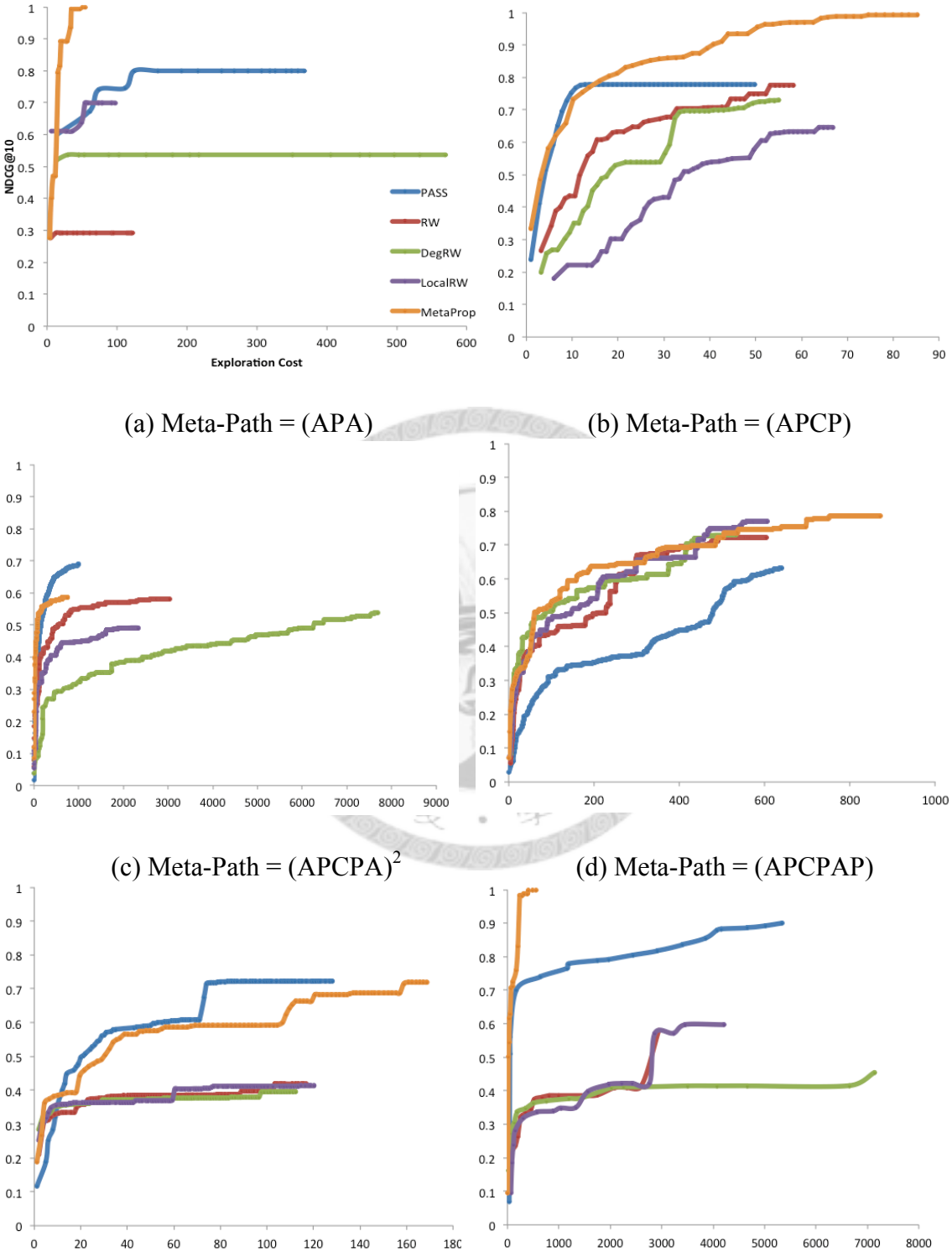
6.2 Evaluating Different Meta-Path Semantics and Datasets

Following the list of meta-path semantics detailed in Table 6-1, we cover a set of distinctive semantics, under different path lengths, to see how the search processes vary due to different semantic objectives. These meta-path semantics have some interesting characteristics. For instance, semantics with shorter meta-path lengths tend to have more proximal connections in a network (such as co-authorship). This can be observed in the experiment results, as the size of the set for sampled objects tend to be small. On

the other hand, longer connections are able to cover more intrinsic explanations that relate two particular objects. We show the results on the High Energy Physics Publication Network for these different semantics in Figure 6-6.

We find that our proposed sampling methodology is able to achieve high NDCG in early sampling stages. In many cases, the proposed method shots up very quickly in the beginning and reaches the top, which remains stable for the remainder of the observation interval. For the other competing methods, NDCG growths are considerably slower. In fact, these curves are more likely to show a staircase-like growth- meaning after some sampling steps, the NDCG values suddenly jump up. This process repeats until stabilization. An intuitive explanation can be viewed in terms of retrieval: given a sampling budget (size of S) and number of items in interest (k in NDCG), we want to have high precision and recall. This means we want to obtain important objects during the sampling by exploration process. More importantly, we want to have high hit rate given the limited budget available, which is nontrivial, but MetaProp and PASS do it easily. However, we observe that there is a variance in NDCG performance for the proposed methods. While MetaProp outperforms baselines throughout different meta-paths, PASS is not always a better performer (though sometimes PASS clearly surpasses other comparing methods, including MetaProp). We again refer to the BFS nature of the sampling algorithm. For the PASS algorithm, if we are traversing on meta-paths with too many repeating node types, it is likely that the sampling algorithm is prone to sample the target node type of interest too early, leading to wrongly interpreted meta-path metric approximation. For example, in APCPAP meta-path, PASS first samples C, a conference node, then upon estimating subsequent paper nodes P, the algorithm would dictate the same probability of selecting P and A (P(type(P)|type(C))=1). In this case, given the high degree of C, the algorithm is going

to sample paper for most of the time. DFS approach, however, avoids fitting local type distribution and topology by traversing meta-path semantic during sampling.



(a) Meta-Path = (APA)

(b) Meta-Path = (APCP)

(c) Meta-Path = $(APCPA)^2$

(d) Meta-Path = (APCPAP)

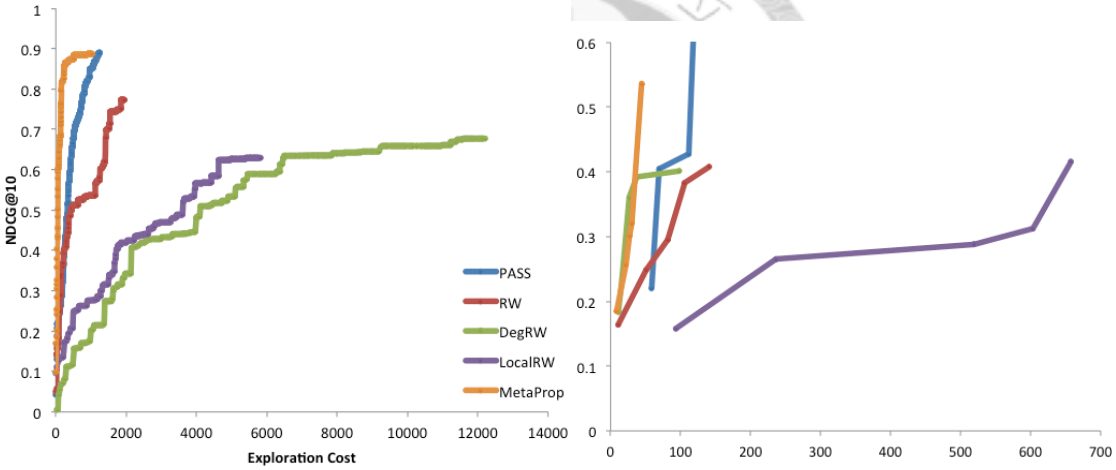
(e) Meta-Path = (CPAP)

(f) Meta-Path = (CPAPC)

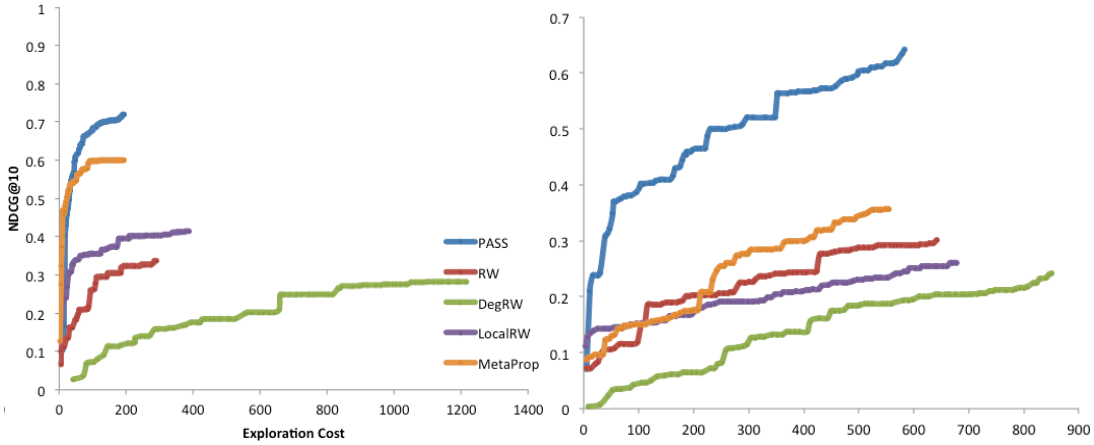
Figure 6-6. NDCG@10 results for different meta-path semantics tested on the High

Energy Physics Publication Network. The measuring metric for the meta-path semantic: PathSim metric for symmetric type and Symmetric Random Walk otherwise.

As already explained earlier, the meta-paths used for the High Energy Physics network bear quite some resemblance to other networks in interest. Therefore, we briefly show the results for the DBLP Publication Network and Movie Network to demonstrate the consistency in the performance of the approach. We illustrate the results in Figure 6-7, using meta-path equivalent= (APCPA), (CPAPC) in HepTh. We find very similar patterns emerging from the two datasets, as the proposed method reaches the top very early. As for the Movie Network, the curves are much more flat, which is likely due to a more complex type relations of the schema.



(a) Results for the DBLP Publication Network. Left: Meta-Path (APCPA) |S| = 500. Right: Meta-Path (CPAPC) |S| = 6



(b) Results for the Movie Network. Left: Meta-Path (APCPA) |S| = 100. Right:

Meta-Path (CPAPC) |S| = 100

Figure 6-7. NDCG@10 results for DBLP Publication Network and Movie Network.

6.3 Examining Different Parameter Settings

Having checked the validity of the proposed sampling method, we turn to examine the setting of the method. We first examine the performance nature in DFS-based algorithms (which are implemented by random surfer algorithms). Acknowledging the random surfer model of the sampling method, first note that calculate distribution on all one-step hidden neighbors and only neighbors of an ego node does not constitute much different in performance. The result reflects the nature of the dataset, as we do not have the fully observed information, so if we attempt to calculate on the whole G_s, there is a good chance of going to local optima.

We look at the assignment of random surfer at restart. Figure 6-8 shows the difference between restarting at the seed node and at random node. We see starting at random node improves accuracy at a later stage, but early retrieval is less effective.

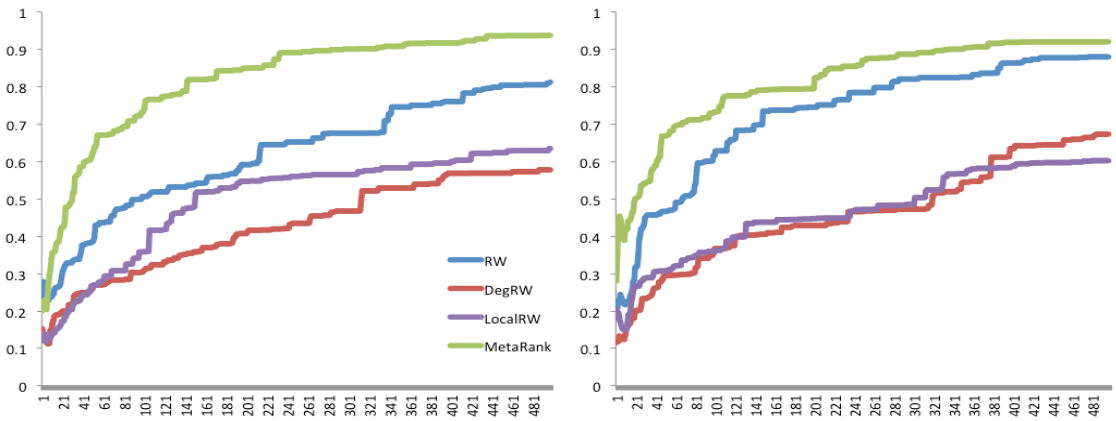


Figure 6-8. NDCG@10 results for High Energy Network. Left: Restarting at Random, Right: Restarting at n_{seed}

We also consider the restart probability for our DFS heuristic, MetaProp. See the diagram in Figure 6-9 for results, which considers High Energy Network's meta-path

(APCPA) at |S| = 250. We see several peaks, and some occur at the endpoints of the [0,1] segment. This presents a tradeoff between pure chain-referral sampling and highly biased neighborhood sampling (that tends to sample in breadth-first search manner). The result suggests careful tuning needs to be considered that account different semantics, network features, sampling budget, and the number of top items of interest.

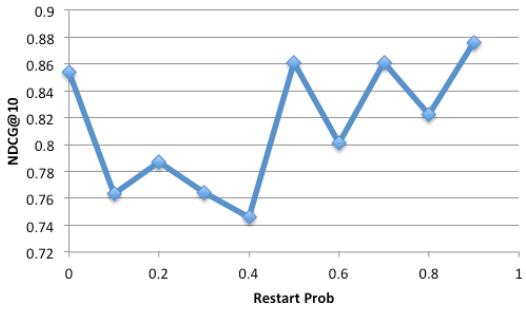


Figure 6-9. Restart probability v.s. NDCG@10 for HepTh, meta-path = (APCPA)

Lastly, we take special interest in checking whether counting paths backwards in PASS would contribute to the metric estimation. To be more exact, when we are computing path completion probability, we add the count table scores from the neighbors of the candidate node. This sum is linearly interpolated with the original estimation term, in which we weight by α. We presume this step should account the bi-directional characteristics exemplified in Pathsim metric. The results for path APCPA in HepTh network are shown in Figure 6-10. We see that the high connecting degree amplifies the local topology, leading to higher cost and lower NDCG.

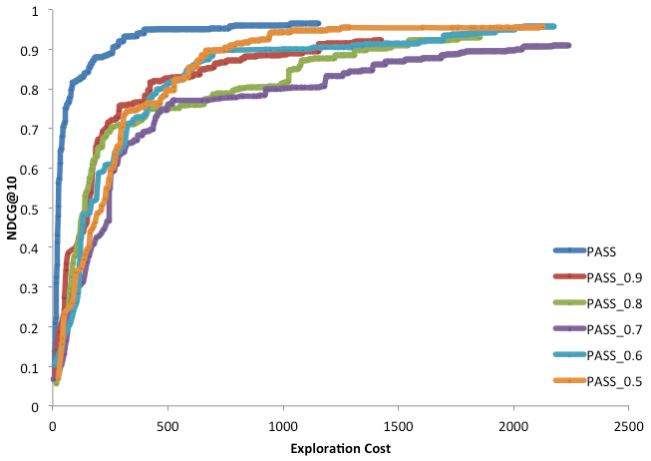


Figure 6-10. NDCG@10 for HepTh, Path=APCPA, for PASS under different α

Chapter 7

Applications

We conclude the experiments with a couple of simple cases to discuss the interpretability of our search results.

7.1 Entity Retrieval Real Examples and Accuracy

Take DBLP Publication Network for instance, we try to retrieve the most relevant conference venues given a query conference and meta-path (CPAPC). We give three exemplifications. Take database domain for example, if we set query = EDBT, which is a premier international database conference, then the sampling results can be found in Table 7-1. The top relevant results to PathSim metric and the sampling methods can be seen in Table 7-1. Results that do not belong to the corresponding domain are labeled in (database:red, data mining:black, IR: orange, machine learning:blue) for reference. Observing the results, we can figure out that the list returned by comparing pure PathSim metric is a list of high profile and specific data mining conferences. We see that Pathsim is quite accurate in terms of interpretation. We further see the result for data mining domain for query = SIAM SDM, and machine learning domain for query = NIPS. Comparing lists by the sampling methods, we realize our proposed method returns highly valued conferences in its early stage. Other methods perform relatively worse in the semantic interpretation. In addition, we find our proposed method return results from consistent domains, even when it returns results from irrelevant domains, whereas other sampling method jumps erratically between different domains.

We also consider correlation with real labels. For the DBLP Publication Network,

the conferences can be classified into four different domains. Given a conference query and meta-path (CPAPC), we would like to return conferences of highly relevant domains. Figure 7-2 shows the result with 7 returned conferences given the query. For this dataset, the highly connected network causes random surfer to easily walk to conferences of different domains. However, in average, the proposed method is able to attain higher accuracy.

Order	RW	DegRW	LocalRW	MetaProp	PASS	PathSim (Ground Truth)
1	ICDT	VLDB	UAI	ICDE	ICDE	ICDE
2	PODS	ICDT	ICDM	CIKM	ICDT	VLDB
3	VLDB	PODS	VLDB	VLDB	PODS	SIGMOD
4	SIGKDD	SIGMOD	SIGMOD	SIGMOD	SIGMOD	CIKM
5	CIKM	CIKM	SIGKDD	ICDT	VLDB	ICDM
6	ICML	WWW	SIGIR	PODS	CIKM	ICDT
7	ECML	ICDE	PODS	WWW	SIGIR	PODS

(a). Query = EDBT, domain = database

Order	RW	DegRW	LocalRW	MetaProp	PASS	PathSim (Ground Truth)
1	ECML	SIGKDD	ICML	ICDM	PKDD	ICDM
2	VLDB	SIGIR	EDBT	SIGKDD	SIGKDD	PKDD
3	CIKM	ICML	WWW	ICML	ICDM	ICDT
4	SIGIR	CIKM	NIPS	NIPS	PAKDD	SIGKDD
5	ECIR	VLDB	SIGKDD	COLT	CIKM	ICML
6	SIGMOD	SIGMOD	PAKDD	CIKM	SIGMOD	PAKDD
7	ICDE	ICDE	ICDM	PKDD	ICDE	CIKM

(b). Query = SIAM SDM, domain = data mining

Order	RW	DegRW	LocalRW	MetaProp	PASS	PathSim (Ground Truth)
1	ICML	SIGIR	ICML	COLT	ICML	COLT
2	PKDD	WWW	SDM	ICML	SIGKDD	ICML
3	WWW	SIGKDD	PAKDD	IJCNN	COLT	UAI
4	UAI	CIKM	IJCNN	UAI	UAI	ECML
5	COLT	SIGMOD	UAI	TREC	ECML	SIGKDD
6	ECIR	VLDB	TREC	SIGIR	CIKM	PAKDD
7	SIGKDD	ICML	SIGIR	ECML	ICDM	ICDM

(c). Query = NIPS, domain = machine learning

Table 7-1. Top Similar Conference given the meta-path semantic (CPAPC) for sampling methods and PathSim score benchmark.

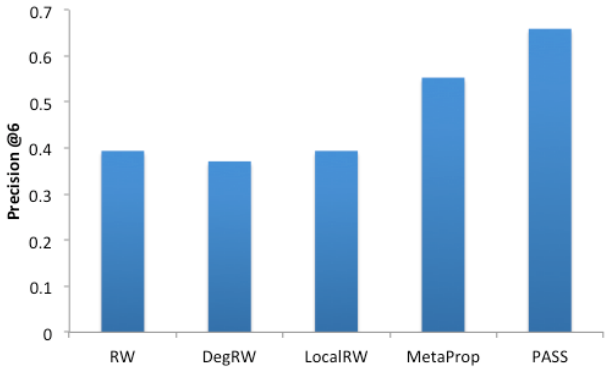


Figure 7-2. Average accuracy for DBLP Network, under meta-path (CPAPC)

7.2 Towards Building Learning to Rank Model with Attribute Information Using Retrieved Target Nodes

We entail another possible user case scenario: in interaction between user program with another application, the other application may provide our user program some information, such as individuals. However, we may not have immediate connection that links the particular individual to the social network. More often, we are only left with some descriptive information of the individual. It is very important to come up with a method to link the individual to the query in consideration and to the network in a

certain way. This is clearly a challenging task, as we do not have any semantic measurement available to provide any sort of estimation.

To cope with this challenge, we show that the set S of nodes with target type T can be used to build learning to rank model to assist prediction of similarity for nodes that do not have any topological information related to the graph but only some content/attributes describing them. Consider the DBLP publication network as an example, where in addition to the currently available three node types (Author, Paper, Conference) we add word attributes for each paper that include the paper's title and abstract (if available). Suppose we are interested in retrieving similar authors to a particular author via meta-path (APCPA). For each author a_i, we represent the author using aggregated information, resulting in $\mathbf{f_{ai}} = [f_{w1}, f_{w2}, \ldots, f_{wk}]$ that means bag of words features with frequency of the word appearing in all a_i's papers as feature value.

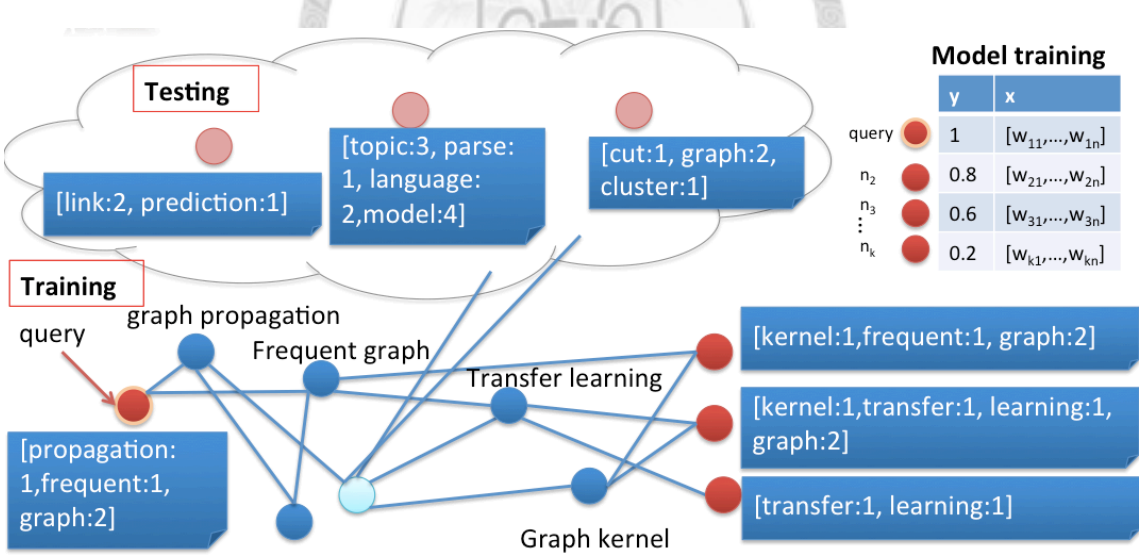


Figure 7-3. Illustration of the application scenario where we wish to predict semantic similarity of nodes to query with only attribute information

Now the problem we are interested can be described as: if we only have an author's attributes (such as $\mathbf{f_{ai}}$) but not how it is connected to the heterogeneous network, can we recover its relative semantic importance to other nodes in the network? To resolve this problem, we propose to use the sampled nodes S to train a ranking model

that can apply to instances with attributes only. An illustration of the process is shown in Figure 7-3.

Specific formulation for the problem goes: we assume as input a set of instances, each derived from a node in S. Suppose each instance i can have features as $\mathbf{f_{ai}}$. In particular, we transform the feature with $\mathbf{f'_{ai}} = \langle \mathbf{f}_{ai}, \mathbf{f}_{n_{seed}} \rangle$, a functional inner product between two feature vectors for instance node a_i and query node n_{seed}. For y's value, we use the currently available meta-path similarity $\mathfrak{M}(a_i, n_{seed})_P$. Using $\mathbf{f'_{ai}}$ as new feature representation for instance i, we build data matrix D, candidate pair matrix P with each pair ($\mathbf{f_a}$, y_a) and ($\mathbf{f_b}$, y_b). To derive prediction function $f(\mathbf{w}, \mathbf{x})$ with $\mathbf{w}$ as model weight function, we apply the combined regression and ranking objective [29]:

$$\min_{\mathrm{w}\in\mathrm{R}^m} \alpha L(\mathbf{w}, \mathrm{D}) + (1-\alpha) L(\mathbf{w}, \mathrm{P}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

Using the above formulation provides a match to our scenario: we are both interested in recovering the meta-path similarity measure, and ensuring the highly similar nodes should be placed in front with highly dissimilar nodes placed far in back of the list, achieving good overall listwise ranking. This goal is exactly what combined regression and ranking objective aims to attain. We use the gradient computation and sampling methods suggested in [29]. For implementation, we use sofia-ml package[7], using pegasos-SVM with default parameters. The ranking objective is tuned to combined regression and ranking.

As for baseline, we consider directly computing the inner product between two bag-of-words vectors for $\mathbf{f}_{ai}, \mathbf{f}_{n_{seed}}$. We compare the result of the baseline against learning to ranking model trained on different sampling methods and different sizes of S.

[7] https://code.google.com/p/sofia-ml/

For the training data, we use the sampled nodes in S to generate features directly. As for testing data, we sample 2000 authors based on the similarity score ranking to the query author. If the number of nodes with similarity greater than 0 is smaller than 2000, we sample the rest of the 2000 authors from the set of all authors that have 0 similarity with the query, and generate the features respectively. Again, we average the result over five different independent query nodes.

To evaluate our result, we take note that our true goal at heart is to retrieve nodes that are highly semantically related, with respect to meta-path semantic measures. Therefore, we measure *precision@k* on the ranked list of predicted scores by the prediction model to see how correct are the retrieved nodes that the model believes to have high semantic similarity scores. To get a rough estimate of k, we apply k-means to divide the instances of ground truth scores into two groups, which can serve as a benchmark that divides the highly dissimilar nodes from the rest of the instances. For the case of this dataset, we use k=300.

See Figure 7-4 for result. We find the naïve baseline performs rather poorly. This is due to highly incomplete information in the instance attributes: some instances have abstracts while others do not, leading to high imbalance of amount of information. Using predicted models significantly improves the performance. In particular, we find the proposed method consistently perform better than the other sampling methods. An important observation is that the precision performance actually increases with more nodes sampled. This corroborates the belief that our proposed method retrieves those nodes with high semantic similarity to the query nodes early in the sampling stage, while other sampling methods may wonder in the network for a long time before finding the nodes that can really boost the predictive model.

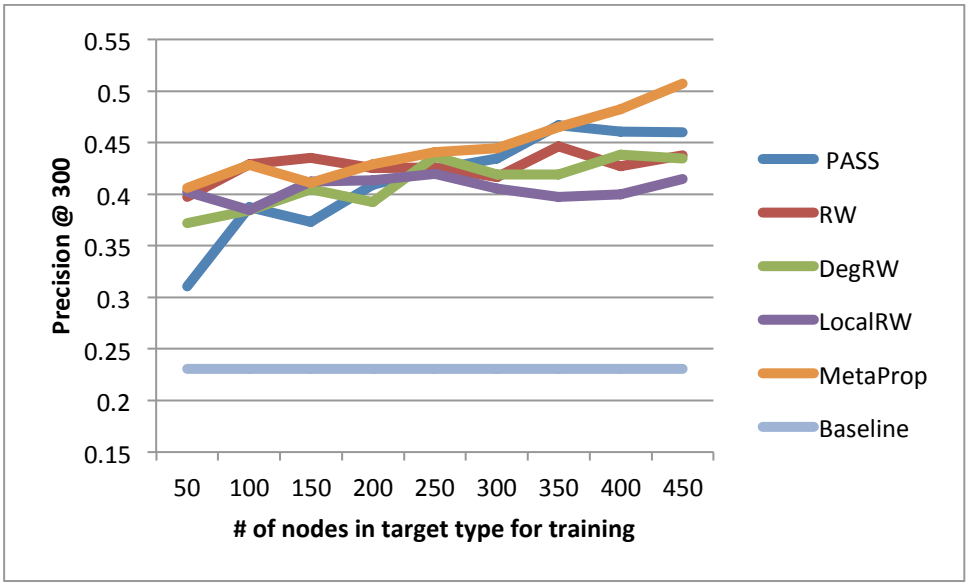


Figure 7-4. Precision@300 versus the number of nodes in S for training the model, comparing using different sampling methods and pure bag-of-words cosine similarity

Chapter 8

Conclusion & Future Work

Heterogeneous social network object search and retrieval using semantics objective is crucial for effective role analysis of each node and how it interacts with other nodes within a graph. In this work, we explore heterogeneous graph entity search based on explorative sampling algorithms. In particular, we provide a generic sampling framework accounting meta-path semantic structure information. A method utilizing probabilistic weighing of candidate nodes is proposed, and the method applies approximation scheme that uses weighted random walk for score computation. We design a series of experiments to verify the validity and usefulness for the sampling algorithm. The experiments confirm our hypothesis of the effectiveness in our sampling algorithm as well as the interpretability of the results. In addition, some application scenarios are introduced and briefly demonstrated in forms of predictive modeling. The future work includes automatically discovering the semantic importance of each meta-path semantic, incorporating a general learning framework for multiple domain attributes into network entity search, and a more general, Bayesian-inference driven prediction model for sampling.

Reference

[1] Gjoka, M. and Butts, C.T. and Kurant, M. and Markopoulou, A,"Walking in Facebook: A Case Study of Unbiased Sampling of OSNs",INFOCOM, 2010 Proceedings IEEE, p.p. 1-9, 2010.

[2] Gjoka, M. and Butts, C.T. and Kurant, M. and Markopoulou, A., "Multigraph Sampling of Online Social Networks", IEEE Journal on Selected Areas in Communications, 29(9):1893 -1905, October 2011.

[3] Heckathorn, Douglas D., "Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations", Social Problems, 49(1), 2002.

[4] Henzinger, M. and Heydon, A. and Mitzenmacher, M. and Najork, M., "On near-uniform URL sampling", Proceedings of the 9th international World Wide Web conference on Computer networks: the international journal of computer and telecommunications networking, pages 295-308, 2000.

[5] Kannan, Ravi and Vempala, Santosh and Vetta, Adrien, "On Clusterings: Good, Bad and Spetral", 2001.

[6] Kasneci, Gjergji and Elbassuoni, Shady and Weikum, Gerhard. "MING: mining informative entity relationship subgraphs". CIKM 2009. 1653-1656.

[7] Leskovec, Jure and Faloutsos, Christos. "Sampling form large graphs", Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 631-636, 2006.

[8] Li, Jhao-Yin and Yeh, Mi-Yen, ”On sampling type distribution from heterogeneous social networks”, Proceedings of the 15th Pacific-Asia conference on Advances in knowledge discovery and data mining - Volume Part II, pages 111-122, 2011.

[9] Lszl Lovsz, ”Random Walks on Graphs: A Survey”, 1993.

[10] Lu, L. and Zhou, T., ”Link Prediction in Complex Networks: A Survey”, CoRR, abs/1010.0, 2010.

[11] Ma, H. and Gustafson, S. and Moitra, A. and Bracewell, D., ”Egocentric Network Sampling in Viral Marketing Applications”, In Proceedings to the 13th IEEE International Conference on Computational Science and Engineering, 2009.

[12] Maiya, A. and Berger-Wolf, T., ”Benefits of Bias: Towards Better Characterization of Network Sampling”, Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011.

[13] Sen, Prithviraj and Namata, Galileo and Bilgic, Mustafa and Getoor , Lise and Gallagher, Brian and Eliassi-rad, Tina, “Collective classification in network data”, 2008

[14] Sun, Y. and Han, J. and Yan, X. and Yu, P.S. and Wu, T., PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. ;In Proceedings of PVLDB. 2011, 992-1003.

[15] Tang, Lei and Wang, Xufei and Liu, Huan, "Uncovering Groups via Heterogeneous Interaction Analysis", Proceedings of the Ninth IEEE International Conference on Data Mining, 2009.

[16] Yang, Cheng-Lun and Kung, Perng-Hwa and Chen, Chun-An and Lin, Shou-De. “Semantically sampling in heterogeneous social networks.” In *Proceedings of the 22nd international conference on World Wide Web companion*,181-182.

[17] Ye, Shaozhi and Lang, Juan and Wu, Felix,"Crawling Online Social Graphs",Proceedings of the 2010 12th International Asia-Pacific Web Conference, pages 236-242, 2010.

[18] Yu, Xiao and Sun, Yizhou and Norick, Brandon and Mao, Tiancheng and Han, Jiawei. "User guided entity similarity search using meta-path selection in heterogeneous information networks." CIKM 2012. 2025-2029.

[19] Yu, Xiao and Sun, Yizhou and Zhao, Peixiang and Han, Jiawei. "Query-driven discovery of semantically similar substructures in heterogeneous networks." SIGKDD 2012. 1500-1503.

[20] Kurant, M. and Gjoka, M. and Butts, C. and Markopoulou, A., "Walking on a graph with a magnifying glass: stratified sampling via weighted random walks," SIGMETRICS, 2011.

[21] Hsieh, Hsun-Ping and Li, Cheng-Te and Lin, Shou-De. "Temporal Social Behavior Search in Heterogeneous Information Networks." WWW 2012.

[22] Yang, Y. and Chawla, N. and Sun, Y. and Han, J., "Predicting Links in Multi-relational and Heterogeneous Networks," ICDM, 2012.

[23] Sun, Yizhou and Norick, Brandon and Han, Jiawei and Yan, Xifeng and Yu, Philip S. and Yu, Xiao. "Integrating meta-path selection with user-guided object clustering in heterogeneous information networks." SIGKDD 2012, 1348-1356.

[24] Navlakha, S. and Rastogi, R. and Shrivastava, N., "Graph summarization with bounded error", Proceedings of the 2008 ACM SIGMOD international conference on Management of dat, pages 419-432, 2008.

[25] Chebolu, Prasad and Melsted, Páll. "PageRank and the random surfer model". In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms* (SODA '08).1010-1018.

[26] Sinatra, Roberta and Gomez-Gardenes, Jesus and Lambiotte, Renaud and Nicosia, Vincenzo and Latora, Vito. “Maximal-entropy random walks in complex networks with limited information”. Physical Review E, vol. 83, Issue 3. 2011.

[27] Li, Rong-Hua and Yu, Jeffrey Xu and Liu, Jianquan. “Link prediction: the power of maximal entropy random walk”. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. 2011, 1147-1156.

[28] Chen, Jiyang and Zaïane, OsmarR. And Goebel, Randy. “Detecting Communities in Social Networks Using Local Information”. From Sociology to Computing in Social Networks, 2010. Page 197-214.

[29] Sculley, D. “Combined Regression and Ranking”. Proceedings of the 16th Annual SIGKDD Conference on Knowledge Discover and Data Mining, 2010

[30] Lichtenwalter, Ryan and Lussier, Jake and Chawla, Nitesh. “*New Perspectives and Methods in Link Prediction*”. KDD 2010