Mining Interesting Meta-Paths from Complex Heterogeneous Information Networks

Baoxu Shi Tim Weninger
Computer Science and Engineering
University of Notre Dame
Notre Dame, Indiana 46556
Email: {bshi, tweninge}@nd.edu

***Abstract*—Meta-paths in heterogeneous information networks are almost always hand created and have, so far, only been attempted on data sets with very small type systems like DBLP, IMDB, etc. Most real-world heterogeneous information networks have large and complex type systems. As the size and complexity of the type-system grows it becomes more and more difficult for humans to form reasonable meta-path queries. This work introduces a new technique to discover a new market for data called *interesting meta-paths* from complex heterogeneous information networks. Our interestingness measure is based on classical knowledge discovery principles, but have been applied in such a way that only interesting meta-paths are mined from the hundreds-of-thousands of possible choices. As in classical pattern mining literature, precision and recall statistics are difficult to obtain; instead we evaluate the effectiveness of our results using a quantitative node-similarity analysis as well as a large user study. Finally, we apply the newly discovered interesting meta-paths to find similar nodes on the Wikipedia heterogeneous information networks.**

***Index Terms*—information networks, meta-paths, similarity**

I. Introduction

Current network science research, for the most part, works with *homogeneous* or *untyped* networks where nodes are objects of the same entity type (*e.g.*, person, protein, particle) and links are relationships of the same type (*e.g.*, friendship, binding, force). This line of research has discovered many influential properties and applications from information networks, including models of contagion in epidemiology [1], [2], small-worlds in social networks [3], [4], power-law distributions on the World Wide Web [5], [6], and so on.

Most real world networks are *heterogeneous*, where nodes and relations consist of different types and have different roles. Heterogeneous information networks (HINs) can be constructed in almost any domain, including social networks (*e.g.*, Facebook), e-commerce (*e.g.*, Amazon and eBay), online movie databases (*e.g.*, IMDB), and in numerous database applications. HINs can also be constructed from text data, such as news collections, by entity and relationship extraction using natural language processing and other advanced techniques.

The heterogeneity of the networks brings rich information but also challenges in the systematic analysis of the connection type between objects. Meta-paths [7] are typed-sequences that connect two or more objects in a HIN. Figures 1 and 2 illustrate a particular path between two Wiki-pages Northeastern University and South Bend, Indiana. These two entries are separated by many loopless paths, one of which traverses through the Wiki-pages of Laszlo Barabasi and the University of Notre Dame. Thus, we can say that Northeastern University is related to South Bend, Indiana via Laszlo Barabasi and the University of Notre Dame. The corresponding meta-path indicated in Figure 2 is ⬠◇⬠□ representing a path of types: Education-People-Education-Geography. This meta-path describes **how** the two endpoints are related. In this particular case, Laszlo Barabasi worked at both Northeastern University and the University of Notre Dame and because the University of Notre Dame is in South Bend, Indiana.

There are many other paths that separate/connect Northeastern University and South Bend, Indiana, and, without loss of generality, each of these paths indicates some special relationship among and between seemingly unrelated objects. In fact, there are 51 alternate paths of length 3 that connect Northeastern University and South Bend, Indiana each indicating some other relationship between the two endpoints. If we include paths of length 4 or 5, then the number of possible paths increases dramatically to thousands and hundreds of thousands. If we further include type-combinations of different granularities for each node, then the problem becomes intractable very quickly.

Current meta-path techniques fail to compute on even moderately-sized networks – recent attempts for HIN analysis needed to limit the popular DBLP dataset to no more than 20 conferences or 1000 authors using only a few, short, hand-crafted meta-paths [8]. Furthermore, the process of creating or hand-annotating meta-paths may be feasible on non-complex type systems like DBLP and IMDB, but hand-annotation is intractable for complex type systems like Wikipedia.

Consider again the Wikipedia example in Figures 1. Because of the complexity of Wikipedia's category/type system, each node contains a hierarchy of types with an increasing granularity as we move up the category-tree. In the corresponding example in Figure 2 we simply chose the first top-level-category that we encountered when performing a breadth-first-search. But is Education the appropriate type-label for Northeastern University? Or should we label it Sports or Ed. in Boston instead?

In this paper we investigate the automatic discovery of *interesting meta-paths* that best describe how two objects

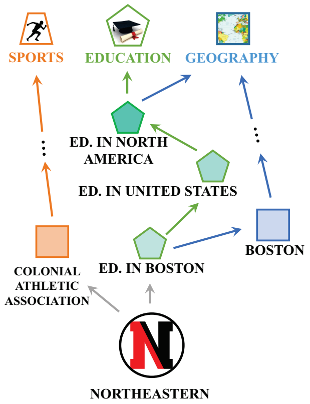

Fig. 1: Type hierarchy on Wikipedia for NORTHEASTERN University

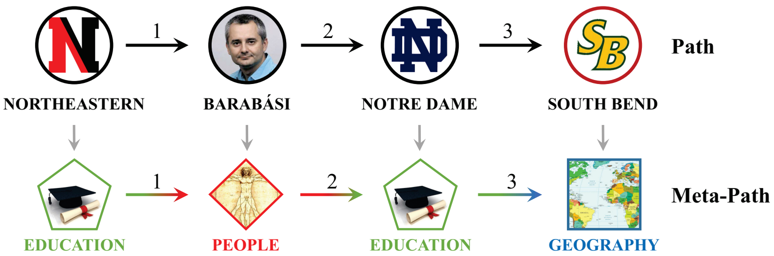

Fig. 2: A path and its corresponding meta-path from NORTHEASTERN University to SOUTH BEND, Indiana. There are many such paths between NORTHEASTERN and SOUTH BEND, that traverse through similar and vastly-different Wiki-pages. We ask: What are the paths that best describe the similarities between two endpoints; what are the best type-labels to assign to each page on this path; and, given type-assignments, can we use the meta-paths to find similar pairs of endpoints?

are uniquely related in complex HINs. Meta-path discovery promises new and interesting insights into the process by which humans collectively organize information: by content similarity, type, temporally, spatially, etc. Unfortunately, there exist many limitations with the current state of the art: 1) Many existing meta-path-based similarity search functions can only find similarities for objects of the same type, that is, given PEOPLE existing methods can only find other PEOPLE [9]; 2) Existing frameworks, such as PathSim, scale only to a few thousand nodes [10], [8]; and 3) Meta-paths must be handcrafted [11], [12].

These challenges will be addressed in this work. Specifically, this paper makes the following contributions:

1) We propose a general framework to mine *interesting paths and meta-paths* from *complex* heterogeneous information networks using adaptations from classical knowledge discovery techniques.
2) We evaluate the robustness of the discovered paths through various large scale experiments on the Wikipedia dataset containing millions of nodes and edges and a complex type system.
3) We explore meta-path discovery at various levels of type-granularity in the complex type systems and discuss the relative trade-offs.
4) We show how the newly discovered meta-paths can be used as a basis for a simple similarity search, and we discuss the implications of type granularity for relative-similarity.

II. RELATED WORK

The task of discovering interesting meta-paths is akin to finding paths in information networks and then analyzing the nature of those paths. There has been a large body of work on pathfinding and similarity in information networks that we will explore, followed by a brief discussion of works that employ meta-paths for other data mining tasks.

If two objects exist in the same network, their similarity can be expressed by the distance between the two objects. If two objects directly reference one another via a graph-edge then they should be considered closely related, tie-strength notwithstanding. If two objects are not directly related, then their similarity can be expressed by some function of the edges and paths that separate them. Studies abound in network science and data mining literature exploring path finding and relatedness, including network navigation [13], [14], decentralized search in networks [3], [15], [16], [17], Web click-trail analysis [18], [19], [20], [21] and so on.

Computing researchers often combine network analysis with text and data mining techniques to validate network theories. For example, the combination of information network analysis of email communication [22], [23] with text analysis of email content can provide a means to test theories such as the strength of weak ties [24], structural holes [25], leadership [26], broadcast diffusion [27], and information navigation [20], [21]. As these studies demonstrate, the examination of information in networks can help researchers and practitioners alike better understand why and how networks play such a vital role in so many physical, real-world phenomena.

The last line of related work can be traced back to Milgrams small-world experiment [3] and the algorithmic problem of decentralized search in networks. Decentralized search considers a scenario in which a starting node s is trying to send a message to a given target node t by forwarding the message to one of its neighbors, where the process continues in the same way until eventually it is reached. This process has been investigated both experimentally as well as through simulations [16], [17].

The critical different between the existing state of the art and the algorithm described in this paper is that existing method require hand-crafted meta-paths to be input by the user, whereas our approach aims to discover the most interesting meta-path from the network.

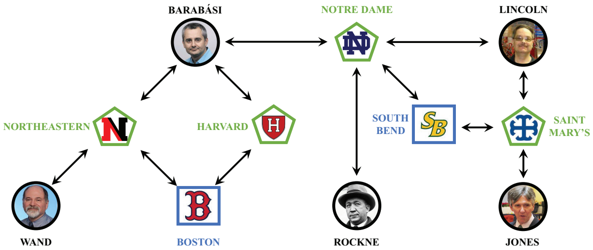

Fig. 3: Heterogeneous Information Network of a simulated social network comprised of People (○), *Geography* (□) and Education (⬠).

III. Interesting Meta-Paths

A. *Simple Network Similarity via Meta-Paths*

We've discussed how the similarity between two objects is related to the size and number of paths that connect them in the network. To distinguish the semantics among paths connecting two objects, meta-paths constrain what paths are explored in the heterogeneous network thereby extracting nodes and paths that have the same types as the meta-path's types. Consider the social network illustration in Figure 3. In this heterogeneous information network there exist three types: People (○), *Geography* (□) and Education (⬠). The persons most similar to Barabasi, considering only graph distance, could be either Wand, Rockne or Lincoln. The different connection scenarios are represented by three distinct meta-paths: (a) ○⬠○, denoting that the similarity is defined by the meta-path "People-Education-People", or (b) ○⬠□⬠○, by the meta-path "People-Education-Geography-Education-People", or (c) ○⬠○⬠○ by the meta-path "People-Education-People-Education-People". A user can choose either (a), (b), (c) or their own combination based on their preferred similarity semantics. According to path (a), Wand and Rockne are equally close to Barabasi because they equally match the meta-path query. According to path (b), Barabasi is equally similar to all of the other Peoples; and according to path (c) Barabasi is again equally similar to all of the other Peoples. The meta-path framework provides a mechanism for a user to select an appropriate similarity semantics by *manually choosing* a proper meta-path.

In most cases the next step is to use that meta-path to determine various separation measures such as path count, normalized path count, random walk probability, symmetric random walk probability, and so on. Those measures are input to some regression, clustering or classification tool for analysis.

B. *Meta-Path Discovery*

As we briefly discussed earlier, we cannot always rely on neatly delineated types that stem from well-structured data sets like DBLP or IMDB. In many cases the network schema consists of a complex ontology or type hierarchy, and in other cases the types in the heterogeneous information network were discovered by imperfect type and role discovery algorithms [28]. In such cases, the network may consist of multiple-membership or hierarchically typed nodes.

We choose Wikipedia because 1) it is one such complex heterogeneous information network, and 2) like DBLP and other popular data sources, Wikipedia has been proved to be a high quality data set [29]. The Wikipedia example illustrated in Figure 1 shows that each Wikipage contains at least one (but usually more than one) category (type), and that each category contains one or more parent categories. Moreover, nothing precludes other untyped information networks such as the World Wide Web or other networks that requiring type and role discovery, from also exhibiting multiple and hierarchical type systems. Thus, it is necessary to develop tools and methodologies to not only cope with these sophisticated systems, but to leverage the complexity for more powerful analytics.

The simplest way to cope with complex type systems is to ignore the hierarchy and heuristically pick some value for each node's type. For example, at the top level in Wikipedia's category system Wikipedia can be organized by its Main Topic Classifications – among other type systems; these include People, Education, Sports, Geography, etc. If each Wikipage is assigned to its nearest Main Topic Classification then a heterogeneous information network emerges.

Paths between individual nodes in the heterogeneous network determine *how* the nodes are separated and/or related. Figure 4 shows some of the short paths between the Wikipedia pages for Mitchell Wand and Knute Rockne. There are actually several hundred paths between Wand and Rockne with at most 4 edges; interestingly, there are three shortest paths of length 3: 1) Wand-Northeastern-Notre Dame-Rockne (path not shown); 2) Wand-Northeastern-Carnegie Mellon University-Rockne; and 3) Wand-Northeastern-Drexel University-Rockne (path not shown). In this case, an obvious question arises: why are Notre Dame, Carnegie Mellon University and Drexel University contained in the short paths while other universities are ignored? Rockne taught and coached at Notre Dame, so that link is obvious. However, upon further investigation we found that the Wiki pages of Drexel University and Carnegie Mellon University mention Rockne in passing because those university's football teams beat Rockne's football team - a feat so notable that it warranted mention on the university's Wikipedia page. Similarly, the path that includes Cy Young and Carl Hubbell is an interesting case: Cy Young has a statue at Northeastern, Carl Hubbell won the Cy Young Award (for excellent pitching), and Rockne is mentioned in a poem along with Carl Hubbell.

Although obscure football victories and poems may make great trivia, they are probably not the best descriptors for the separation and/or relatedness between Wand and Rockne considering the plethora of alternatives. Again the challenge arises: **how do we determine the most interesting paths that relate and separate two nodes?** This is a clear data mining task that asks for an algorithmic solution.

With these examples in mind, we developed an algorithm to mine *interesting* meta-paths from heterogeneous information networks.

Because a network path is basically a sequence of items/nodes, it is possible to turn to sequential pattern mining literature for help in developing appropriate interestingness measures. In order to mine interesting network-paths using the sequential pattern mining paradigm we could create a database of possible paths between the two nodes and then run sequential pattern mining algorithms over the path-database. Unfortunately, the number of possible paths that would populate the sequential database is intractable, even in relatively small networks and even under reasonable path-length constraints.

C. Path Generation

The proposed path generation technique is akin to a generate-and-discard approach that first generates a set of paths between two objects and throws away uninteresting paths. Figure 5 shows an intuitive example of the proposed path mining method given two query nodes WAND and ROCKNE. First we find candidate paths and nodes by collecting the paths between WAND and ROCKNE. Second we consider all of the type or role-siblings of WAND and ROCKNE and find paths between each. For example, WAND and BARBARA LISKOV are both type-siblings in this example because they share a Wiki-category called AMERICAN COMPUTER SCIENTISTS. The candidate paths are collected into a list X and the sibling-paths are collected into a list Y. By comparing X and Y with respect to some interestingness measure we will be able to discover the candidate paths that are most *interesting* and worth returning to the user.

For this work we require a start-point a_u and endpoint a_v to be provided by the user. These inputs determine the two objects that should be analyzed. Using these two endpoints we find 100 short paths that separate a_u and a_v using Yen's k-shortest paths algorithm [30] where $k = 100$. As a result we have k paths of variable (yet mostly short) length: $\vec{x} = \langle a_1, a_2, \ldots, a_{|\vec{x}|}\rangle_i$ s.t. $a_u = a_1, a_v = a_{|\vec{x}|}, 1 \leq i \leq k$.

Recall from the example in Figure 1 that each node has one or more parent-types $\{t_1, t_2, \ldots\}$, and inversely, each type will contain one or more nodes $\{a_1, a_2, \ldots\} \in t$ that are of an equivalent type. For a given node a_i, we define a_j to be a sibling of a_i if they share the one or more types, *i.e.*, sib(a_i, a_j) iff $a_i \in t \wedge a_j \in t$. Furthermore, for a given node a_u there are many nodes that satisfy the above sibling-definition; the set of siblings of a_u is $\{a'_{u_1}, a'_{u_2}, \ldots\} \in A'_u$.

In order to determine which path $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_k\} \in X$ is the most interesting path we compare it with sibling paths $\{\vec{y}_1, \vec{y}_2, \ldots\} \in Y$. Sibling paths are generated with a modified version of Yen's k-shortest path algorithm that finds k paths between all start-point a_u's siblings A'_u and all of the endpoint a_v's siblings A'_v. Figure 5 illustrates the make up of the short-paths X and the short-sibling-paths Y.

Note that this process generates *interesting paths* not meta-paths. However, we shall see that an exciting side-effect of these calculations is the emergence of one or more **interesting meta-paths**.

Next we describe two interestingness calculations. The first uses an unordered collection of all the types on the short-paths. The second looks at the node types in sequence.

D. Unordered Analysis

Each type t can have one or more parent types themselves creating a hierarchy of types. For the unordered case, T_a represents the full set of parent-types and all ancestor-types for a node a. Note that the type granularity or hierarchy is not considered in this set-of-types representation.

If we apply this to a path of nodes $\vec{x}_i = \langle a_1, a_2, \ldots, a_{|\vec{x}_i|}\rangle$ we can retrieve a path of sets-of-types $\vec{T}_{\vec{x}_i} = \langle T_{a_1}, T_{a_2}, \ldots, T_{a_{|\vec{x}_i|}}\rangle$, or we can simply combine all of the sets-of-types on the path into a single set-of-set-of-types: $T_{\vec{x}_i} = \bigcup_{n=1}^{|\vec{x}_n|} T_{a_n}$.

If we similarly apply this to a sibling path $\vec{y}_i = \langle a'_u, \ldots, a'_v\rangle$ we can get a path of sets-of-types for the sibling paths $\vec{T}_{\vec{y}_i} = \langle T_{a'_u}, \ldots, T_{a'_v}\rangle$. We can again simply combine all of the sets-of-types on the path into a single set-of-set-of-types $T_{\vec{y}_i} = T_{a'_u} \cup \bigcup_{n=2}^{|\vec{y}_i|-1}\{T_{a_n}\} \cup T_{a'_v}$. We want to discover interesting meta-paths using all of the types in all of the sibling paths, so we further add all of the paths together to get one large set of all of the types in all of the elements in all of the sibling paths: $T_Y = \bigcup_{i=1}^{|Y|}\{T_{\vec{y}_i}\}$.

Finally, we define the rank of each path according to the proportion of types that appear in the path but not in the set of sibling paths:

$$r(\vec{x}_i) = \frac{|T_{\vec{x}_i} \cap T_Y|}{|T_Y|}$$

E. Ordered Analysis

By taking the ordering of the nodes on the paths into account we are able to specifically account for differences in node (and node-type) positions. To do this we again consider a set of k-shortest paths $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_k\}$ between start-point a_u and endpoint a_v, and the corresponding sibling paths $\{\vec{y}_1, \vec{y}_2, \ldots\} \in Y$. We again consider the type parents for each node T_a, but this time we consider the set-of-types one position at a time, such that T_{a_1} is not combined with T_{a_2}, etc. Instead, we look at the type proportions at each path position:

$$p(a_n, a'_n) = \frac{|T_{a_n} \cap T_{Y_n}|}{|T_{Y_n}|}$$

where T_{Y_n} is the union of the set-of-types for each node in the n^{th} position on the sibling paths: $\bigcup_{i=1}^{|Y|}\{T_{a_{i,n}}\}$.

To find the final proportion under ordered-path conditions we multiply the positional proportions together:

$$r(\vec{x}_i) = \prod_{n=1}^{|\vec{x}_i|} p(a_n, a'_n)$$

Because $0 \leq p(a_n, a'_n) \leq 1$, longer paths will be at a disadvantage. So, we could alternatively calculate a path's rank based on the mean proportion.

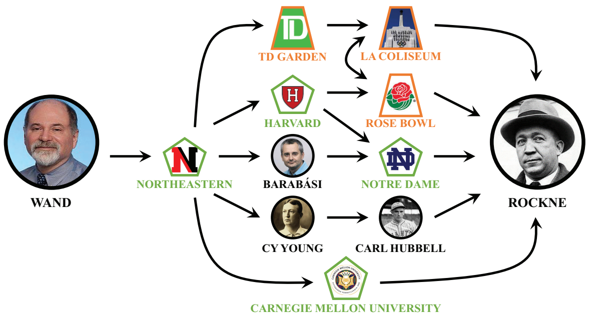

Fig. 4: Small sample of short paths between Mitchell WAND and Knute ROCKNE. WAND is currently a professor at NORTHEASTERN, and ROCKNE was a student, instructor and coach at NOTRE DAME. We aim to find the path that best describes how WAND and ROCKNE are related/separated. Once found, this *interesting meta-path* will be used to find other nodes that are related to WAND in the same way that ROCKNE is related to WAND.

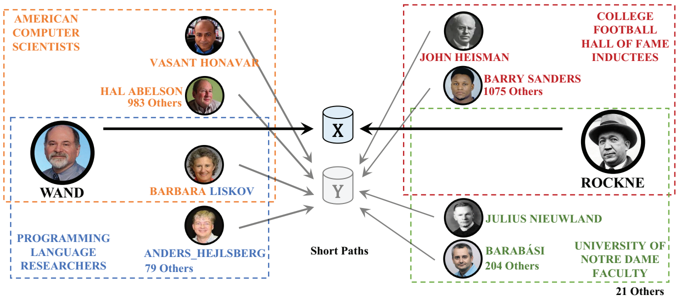

Fig. 5: Example of proposed interesting path mining method. First, find short paths X between input/query nodes WAND and ROCKNE. Second, find short paths Y between the type or role-siblings of WAND and ROCKNE. Finally, use the paths in Y to determine which paths in X are interesting.

In the example from Figure 5 we may find that an interesting path is WAND-NORTHEASTERN-HARVARD-ROSE BOWL-ROCKNE because ICE HOCKEY is found to separate WAND and ROCKNE, but not their type-siblings. We would also find that a meta-path between WAND and ROCKNE would be PEOPLE-EDUCATION-EDUCATION-SPORTS-PEOPLE, or, more-probably, a more fine-grained version of this meta-path that we'll discuss later.

F. Discussion

The unordered analysis is very similar to the confidence measure in classical association rule mining *c.f.*, $p(A \cap B)/p(B)$ [31], and the ordered analysis can be thought of as a product of multiple, smaller confidence measures. Of course, any number of interestingness measures can be applied in various different ways.

What do these rankings mean? The ordered and unordered analysis presented here describe how the paths by which the start point and the endpoint are connected with respect to their types. A path $\vec{x}_i = a_u \rightsquigarrow a_v$ that does *not* share many types in common with the sibling paths would receive a score $r(\vec{x}_i)$ closer to 0 than a path with many types in common with the sibling paths. For lack of a better terminology, the $\vec{x}_i$ with the score closest to 0 is called the most interesting path between a_u and a_v. In other words, the most interesting path is the path (from within the k-shortest paths between a_u and a_v) that most uniquely separate a_u from a_v.

Conversely, the path with the score closest to 1 is called the most-general or least interesting path between a_u from a_v. The most general path is so called because it has the most types in common with the types in the set of sibling paths.

Examples and analysis of interesting and general paths are presented in the next section.

IV. PATH RANKING ANALYSIS

To concretely evaluate the utility of the proposed model we have devised a quantitative experiment and two qualitative experiments. The first qualitative evaluation uses several thousand human judgements to make sense of the variously ranked paths; the second qualitative evaluation illustrates typical and extreme cases[1]. Although there exist other heterogeneous information network similarity measures as discussed above, none of the existing techniques are comparable because they require hand-crafted meta-paths as input. Conversely, the goal of this paper is to determine the important/interesting meta-paths.

A. Dataset

We use a very large, heterogeneous, directed network datasets in our evaluation: Wikipedia (from the Dec. 2, 2013 database dump). We chose Wikipedia because of their robust type-system and their size. DBLP, IMDB and other commonly used information network datasets could be used, but their limited type systems are not the focus of meta path discovery and therefore would not appropriately demonstrate the robustness of the proposed model.

TABLE I: Experimental Dataset consisting of large, complex heterogeneous information networks

	nodes	edges	types	avgdl
Wikipedia	10,276,554	740,056,056	1,018,609	524.7

Table I shows the sizes of the dataset. These statistics includes redirection nodes that are not filtered out during processing. Avgdl is the mean document length in the number of words per document/node. Edges correspond to Wiki-links

[1] The entire set of ranked, result paths as well as source code and raw data are available at https://github.com/nddsg/discrmetapath

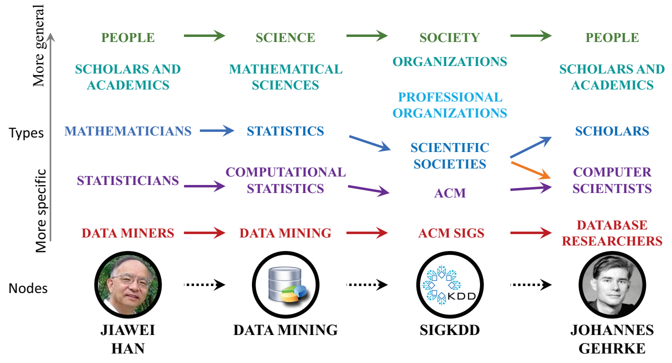

Fig. 6: Example of interesting paths from Wikipedia graph. Text above the nodes represent portions of the type hierarchy for each node. Colored arrows connecting types correspond to meta-paths at various granularity levels. The Wikipedia paths are expanded and illustrated in Figure 7. (This figure is best viewed in color.)

in Wikipedia. The number of types in Wikipedia corresponds to the number of Wikipedia Categories and is discussed throughout this paper.

To generate random input for the start-node a_u and the end-node a_v we randomly picked a starting node a_u and then repeatedly picked a candidate ending-nodes until a node was found to exist within 4 hops of the starting node. This was done to limit the computational complexity of dealing with very long paths. Furthermore, we are not particularly interested in overly long paths with looking for interesting meta-paths. Overall we generated 2,979 paths from Wikipedia.

Each pair of random inputs (a_u, a_v) is considered by the path generation algorithm outlined above and ranked. We report the paths at the top (most interesting), one-quarter, half, three-quarter and bottom (least interesting) of the rankings, labeled 0, 0.25, 0.5, 0.75, 1.0 respectively throughout. This allows us to determine what differences, if any, exist between the path rankings.

V. Meta-Path Selection

Once we obtain interesting paths through the graph, the next step is to find the corresponding meta-path. For example, the path from Wand-Northeastern-Harvard-Rose Bowl-Rockne would have several possible meta-paths including People-Education-Education-Sports-People corresponding to the main-topic classifications, *i.e.*, top-level categories, on Wikipedia. Of course, it may also be beneficial to consider the lowest category-granularity resulting in a meta-path such as Programming Language Researcher-Education in Boston-Association of American Universities-Rose Bowl-College Football Hall of Fame Inductees which correspond to the one of immediate category designations for each node.

Recall, as illustrated in Figure 1, that each node can have multiple category designations and each category can themselves have multiple parent-categories. For the remainder of this paper we simplify the category tree into a category vector by finding the shortest path from the node to the top-level category. Taking the category hierarchy in Figure 1 as an example, we would simplify Northeastern's category hierarchy to be a vector of Education in Boston-Education in the United States-Education in North America-Education, because the path to the top-level Education category was the shortest. Ties are broken arbitrarily.

Note that this simplification is for clarity on the Wikipedia data; simplified category hierarchies are not central to our approach.

Next, we change our running example to a pair of computer scientists Jiawei Han and Johannes Gehrke. Figure 6 illustrates that the most interesting path passes through the Data Mining and SIGKDD nodes[2]. Above each node is the simplified category vector with the most specific category on the bottom and the most general category on the top.

Depending on our task, we may wish to choose meta-paths at different granularity levels. We define a granularity parameter λ that regulates the specificity of a meta-path, where $0 \leq \lambda \leq 1$, and 0 means most specific and 1 means most general. For example, $\lambda = 0$ would return a fine-grained meta-path best suited for the discovery of specific Database Researchers that are linked to by ACM SIG nodes, and so on (red arrows in Figure 6). A $\lambda = 1$ would return a meta-path of the top-level categories (green arrows in Figure 6).

VI. Simple Meta-Path Similarity

In this section we present results for a simple meta-path similarity search on Wikipedia. As is common in heterogeneous information network literature, no standard evaluation procedure exists for this type of analysis, so in lieu of precision and recall scores we present the reader with results from different points of view.

Specifically, we take the interesting meta-paths at various λ-granularity and perform a constrained random walk with restart (RWR) search with 1,000 iterations. When we reach the end of the constraining meta-path, the corresponding node is recorded as having been visited. Similarity between the visited nodes and the query nodes is obtained by ranking visited nodes by their visitation percentage. For example, if a single meta-path-constrained RWR starting at Jiawei Han ends at Johannes Gehrke then the count for Johannes Gehrke is increased by one. The percentage of constrained walks ending with Johannes Gehrke out of all completed walks is the similarity score.

More robust and complete clustering and ranking algorithms exist in recent literature. These heterogeneous information network clustering and ranking algorithms operate, at a fundamental level, by looking at various meta-path-constrained random walks. Rather than including the newly discovered interesting meta-paths as features in the existing algorithms, we are instead interested in showing the basic properties and behavior of the interesting meta-paths.

[2] ICDM, unfortunately, does not have a Wikipedia page

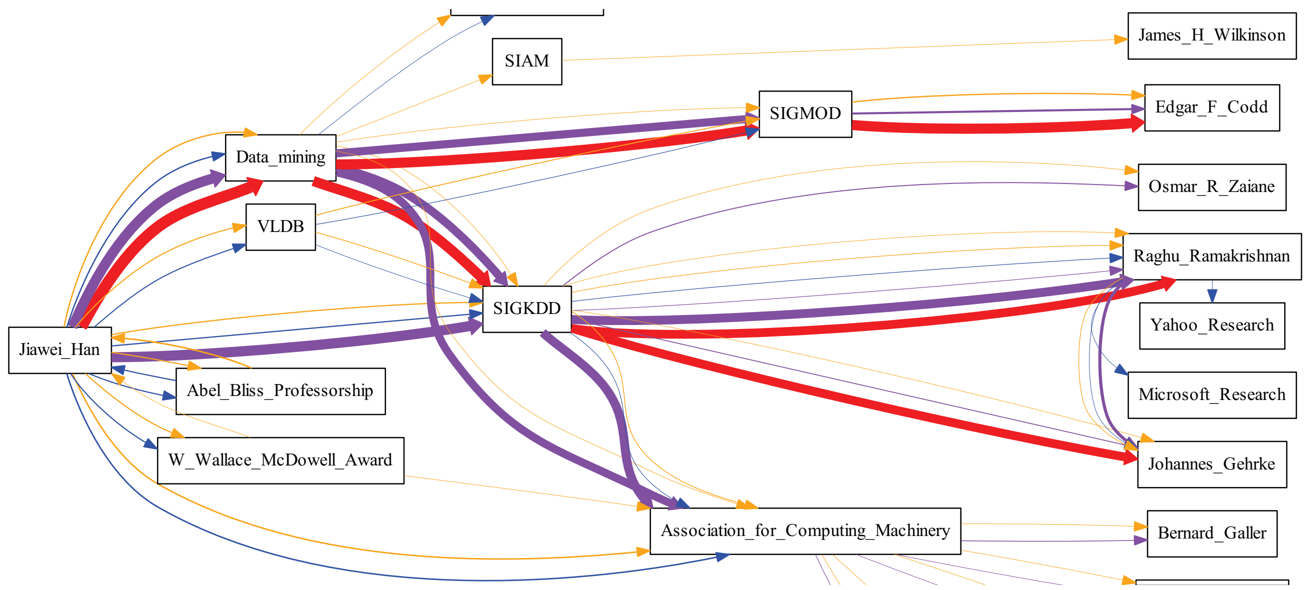

Fig. 7: Illustration-graph of the most frequent paths traveled with constrained RWR with meta-paths of different granularity *i.e.*, λ values. The colors of the paths in Figure 7 correspond to the colors of the meta-paths in Figure 6. The line-thickness corresponds to the probability of traversing a given edge; edges under 10% probability were not drawn for clarity. (This figure is best viewed in color.)

TABLE II: Results of meta-path constrained RWR for various λ values on Wikipedia. Blank values mean that the node was never encountered in 1,000 random walks.

	λ			
	0	0.24	0.41	0.48
Edgar F. Codd	40.5	18.1	9.0	
Johannes Gehrke	28.4	29.4	8.4	2.8
Raghu Ramakrishnan	31.1	6.0	3.6	
Anita Borg		5.1	0.6	0.2
Shafi Goldwasser		4.9	0.6	
Osmar R. Zaiane		4.8	3.6	1.6
Vint Cerf		4.1	2.4	0.2
Allen Newell		2.0	0.6	
ACM				5.1
IEEE				4.9
Yahoo! Research				4.8
Microsoft Research				4.4

Table II shows the probability of reaching an endpoint using the constrained RWR at different levels of λ. The meta-paths generated by the λ values in Table II correspond to the the meta-paths shown in Figure 6. The bottom-most meta-path DATA MINERS-DATA MINING-ACM SIGS-DATABASE RESEARCHERS corresponds to $\lambda = 0$ meaning it is the meta path at the finest granularity. The top meta-path of PEOPLE-SCIENCE-SOCIETY-PEOPLE returned an extremely wide array of PEOPLE, none more than 3 times (out of the thousand RWR iterations), so results for $\lambda = 1$ were omitted in Table II.

Recall that these similarity results listed here are not necessarily the people or nodes most similar to JOHANNES GEHRKE. Rather they are the nodes that are similar to JOHANNES GEHRKE *in the same ways* that JIAWEI HAN is similar to JOHANNES GEHRKE. A different starting point would surely return a different set of results.

As expected, different meta-path granularities give different types of results. Presumably this is because different type granularities cast a wider net of possible nodes. In other words, general types impose less of a constraint on the random walker than specific, or fine-grained, types.

A. *Exploring Interesting Meta-paths*

The previous section presented results nodes that were found to be related with a target node in the same way that the source node is related to the target node. The presented results were gathered with a very simple similarity algorithm – meta-path constrained RWR. The state of the art in HIN clustering and classification use several different network measures to generate values that are input into a regression, clustering or classification algorithm. We omitted those steps, and instead show the raw features. The actual presented results are less important than the understanding of how the interesting meta-paths actually describe the relationships between the start points and endpoints.

To that end we traced the paths followed by the constrained RWR algorithm for JIAWEI HAN and JOHANNES GEHRKE. Figure 7 shows the most frequent paths traveled under different λ values. The colors of the paths in Figure 7 correspond to the colors of the meta-paths in Figure 6. The line-thickness corresponds to the probability of traversing a given edge; edges under 10% probability were not drawn for clarity.

The nodes listed on the right-hand portion of the illustration correspond to the nodes most similar to JOHANNES GEHRKE under the different meta-path granularities. For example, YAHOO RESEARCH and MICROSOFT RESEARCH are found to be related to JOHANNES GEHRKE, but *only* at the most-course granularity setting; this is because YAHOO RESEARCH and MICROSOFT RESEARCH are within the SCHOLARS-type. Otherwise, the endpoints are COMPUTER SCIENTISTS or DATABASE RESEARCHERS or both.

Also note that the Wikipedia Category DATABASE RESEARCHERS contains 48 total Wiki-pages corresponding to many well known database researchers that are presumably, somehow, related to JOHANNES GEHRKE. However, we stress that, in this case, the 47 other database researchers have been found to not be related to JOHANNES GEHRKE in the same "interesting" way that JIAWEI HAN is related to JOHANNES GEHRKE. Thus many of the 47 are not included in the results.

VII. CONCLUSIONS

In conclusion, we have presented an algorithm that discovers interesting paths from complex heterogeneous information networks. Next we performed an analysis of the paths at various levels of interestingness and found no statistically significant difference between the similarities of the documents on the paths. We interpret this negative result to mean that the differences in types do not necessarily correspond to a significant difference in the overall word distribution. A qualitative analysis on the same paths found that humans judges chose the path with the highest interestingness score as the path that *best* separates two random nodes.

Next we showed how meta-paths of varying granularity can be extracted from the interesting paths and used to find nodes that are similar to a given endpoint in the same way that the starting point is similar to the endpoint. Finally, we presented a brief snapshot of the paths that were traversed during the meta-path constrained RWR.

As a matter for future work we intend to use human paths to inform the selection of interesting meta-paths. We will also explore the potential that a generative model could have to iteratively refine other interesting meta-paths.

In summary, the complex type systems that exist in large, real-world heterogeneous information networks pose a problem for existing techniques. We have presented a way to algorithmically present the user with interesting meta-paths that can be used to issue incisive queries to complex heterogeneous information networks.

ACKNOWLEDGMENT

We thank Maria Glenski for her help in preparing this manuscript.

REFERENCES

[1] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease." *Nature reviews. Genetics*, vol. 12, no. 1, pp. 56–68, Jan. 2011.

[2] W. O. Kermack and A. G. McKendrick, "A Contribution to the Mathematical Theory of Epidemics," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 115, no. 772, pp. 700–721, 1927.

[3] S. Milgram, "The small world problem," *Psychology Today*, vol. 1, no. 1, pp. 60–67, 1967.

[4] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks." *Nature*, vol. 393, no. 6684, pp. 440–2, Jun. 1998.

[5] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the Internet topology," *ACM SIGCOMM Computer Communication Review*, vol. 29, no. 4, pp. 251–262, Oct. 1999.

[6] L. A. Adamic, B. A. Huberman, A.-L. Barabási, R. Albert, H. Jeong, and G. Bianconi, "Power-Law Distribution of the World Wide Web," *Science*, vol. 287, no. 5461, p. 2115, Mar. 2000.

[7] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," in *PVLDB*, 2011.

[8] 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," *ACM Transactions on Knowledge Discovery from Data*, vol. 7, no. 3, p. 11, Sep. 2013.

[9] Y. Sun, C. C. Aggarwal, and J. Han, "Relation strength-aware clustering of heterogeneous information networks with incomplete attributes," *PVLDB*, vol. 5, no. 5, pp. 394–405, Jan. 2012.

[10] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, "Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks," in *ASONAM*, 2011.

[11] N. Lao and W. W. Cohen, "Fast Query Execution for Retrieval Models based on Path Constrained Random Walks," in *SIGKDD*. New York, New York, USA: ACM Press, 2010.

[12] C. Shi, X. Kong, Y. Huang, P. S. Yu, and B. Wu, "HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks," in *EDBT*, New York, New York, USA, 2012.

[13] J. Muramatsu and W. Pratt, "Transparent Queries," in *SIGIR*. New York, New York, USA: ACM Press, Sep. 2001, pp. 217–224.

[14] C. Olston and E. H. Chi, "ScentTrails," *ACM Transactions on Computer-Human Interaction*, vol. 10, no. 3, pp. 177–197, Sep. 2003.

[15] O. Simek and D. Jensen, "Navigating networks by using homophily and degree." *PNAS*, vol. 105, no. 35, pp. 12 758–62, Sep. 2008.

[16] L. A. Adamic and E. Adar, "How to search a social network," *Social Networks*, vol. 27, no. 3, pp. 187–203, 2005.

[17] J. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, no. 6798, p. 845, Aug. 2000.

[18] M. Bilenko and R. W. White, "Mining the search trails of surfing crowds," in *WWW*. New York, New York, USA: ACM Press, Apr. 2008, p. 51.

[19] E. Adar, J. Teevan, and S. T. Dumais, "Large scale analysis of web revisitation patterns," in *CHI*. New York, New York, USA: ACM Press, Apr. 2008, p. 1197.

[20] R. West and J. Leskovec, "Human wayfinding in information networks," in *WWW*. New York, New York, USA: ACM Press, Apr. 2012, p. 619.

[21] R. West, J. Pineau, and D. Precup, "Wikispeedia: an online game for inferring semantic distances between concepts," in *IJCAI*. Morgan Kaufmann Publishers Inc., Jul. 2009, pp. 1598–1603.

[22] B. Klimt and Y. Yang, "The Enron Dataset: A New Dataset for Email Classification Research," in *ECML*, ser. Lecture Notes in Computer Science, J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Eds., vol. 3201. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 217–226.

[23] A. McCallum, X. Wang, and A. Corrada-Emmanuel, "Topic and role discovery in social networks with experiments on enron and academic email," *Journal of Artificial Intelligence Research*, vol. 30, no. 1, pp. 249–272, Sep. 2007.

[24] M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[25] R. S. Burt, "Structural Holes and Good Ideas," *American Journal of Sociology*, vol. 110, no. 2, pp. 349–399, Sep. 2004.

[26] M. Zimmermann and V. Eguíluz, "Cooperation, social networks, and the emergence of leadership in a prisoners dilemma with adaptive local interactions," *Physical Review E*, vol. 72, no. 5, p. 056118, Nov. 2005.

[27] M. Gomez Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *SIGKDD*. New York, New York, USA: ACM Press, Jul. 2010, p. 1019.

[28] T. Weninger, C. Zhai, and J. Han, "Building enriched web page representations using link paths," in *HT*. New York, New York, USA: ACM Press, Jun. 2012, p. 53.

[29] J. Giles, "Internet encyclopaedias go head to head," *Nature*, vol. 438, no. 7070, pp. 900–901, 2005.

[30] J. Y. Yen, "Finding the k shortest loopless paths in a network," *management Science*, vol. 17, no. 11, pp. 712–716, 1971.

[31] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," in *VLDB*, 1994.