Multi-Relational Link Prediction in Heterogeneous Information Networks

Darcy Davis, Ryan Lichtenwalter, Nitesh V. Chawla
Interdisciplinary Center for Network Science and Applications
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN, 46556 US
{ddavis4,rlichten,nchawla}@nd.edu

***Abstract*—Many important real-world systems, modeled naturally as complex networks, have heterogeneous interactions and complicated dependency structures. Link prediction in such networks must model the influences between heterogenous relationships and distinguish the formation mechanisms of each link type, a task which is beyond the simple topological features commonly used to score potential links. In this paper, we introduce a novel probabilistically weighted extension of the Adamic/Adar measure for heterogenous information networks, which we use to demonstrate the potential benefits of diverse evidence, particularly in cases where homogeneous relationships are very sparse. However, we also expose some fundamental flaws of traditional *a priori* link prediction. In accordance with previous research on homogeneous networks, we further demonstrate that a supervised approach to link prediction can enhance performance and is easily extended to the heterogeneous case. Finally, we present results on three diverse, real-world heterogeneous information networks and discuss the trends and tradeoffs of supervised and unsupervised link prediction in a multi-relational setting.**

I. Introduction

In network science, the link prediction task can be broadly generalized as follows: Given disjoint source node s and target node t, predict if the node pair has a relationship, or in the case of dynamic interactions, will form one in the near future [1]. For many real world scenarios, link prediction can be applied to anticipate future behavior or to identify probable relationships that are difficult or expensive to observe directly. In social networks, link prediction can be used to predict relationships that will form, uncover relationships that probably exist but have not been observed, or even to assist individuals in forming new connections [2]. In biomedicine, where exhaustive, reliable experimentation is usually not viable, link prediction techniques such as disease-gene candidate detection are valuable for navigating incomplete data, as well as guiding lab resources toward the most probable interactions.

Many interesting real-world systems form complex networks with multiple distinct types of inter-related objects and relationships. Structures of this type are broadly defined as heterogeneous information networks [3], an umbrella term which encapsulates multi-mode, multi-dimensional, multi-relational, and bipartite networks [4]. Link prediction is these networks has typically been performed by treating all relationships equally or by separately studying homogeneous projections of the networks and ignoring dependency patterns across types. Both of these approaches represent a loss of information. Different edge types may have a different topology or link formation mechanisms but nonetheless influence each other, such that various combinations have different relevance to the link prediction task. For example, heterogenous relationships such as friendship, family, and colleague are often modeled as indistinct in social networks. In reality, though, it may be far more probable for person to form a new interaction with the colleague of a colleague than with the mother of a colleague. In the biological domain, networks are often modeled from a singular view of the cell, such as physical protein interaction only. Since many current molecular data sets are unreliable and many interaction types are correlated, it seems natural that integrating diverse information would be mutually beneficial. There is just one problem: the lack of an effective, general methods for link prediction in heterogeneous information networks.

In this paper, we describe and evaluate two new approaches to the link prediction task in heterogeneous information networks. In Section II, we describe three real-world heterogeneous data sources and our evaluation framework. In Section III, we provide a brief survey of standard link prediction methods. We then propose a probabilistic weighting scheme for extending the Adamic/Adar measure to heterogeneous data in Section III-C. Our experiments demonstrate that this extension can effectively use diverse evidence to improve performance over Adamic/Adar in many cases. On the other hand, our results also highlight the fact that extending existing methods cannot overcome the domain-specificity of unsupervised link predictors. In Section IV, we show that, just like homogeneous problems [5], supervised link prediction is still the best available choice in terms of performance. This fundamental change to a general, variance-controlled, data-driven model is even more important in heterogeneous networks, where multiple distributions and decision boundaries are more of a guarantee than a possibility. Finally, we discuss interesting observations and conclusions in Section V.

II. Data

We use three real-world heterogeneous information networks to demonstrate the methods in this paper. The networks

978-0-7695-4375-8/11 $26.00 © 2011 IEEE
DOI 10.1109/ASONAM.2011.107

were chosen from different domains and are considerably divergent in structure and relationship types, which will support the generality of our conclusions.

A. YouTube Network

The YouTube network is constructed from data crawled from the popular video sharing site in December 2008. The crawl collected information about contacts, favorite videos, and subscriptions. In total, it reached 848,003 users, with 15,088 users sharing all of the information types. These 15,088 users are the nodes in the social network, connected by a network of five different interaction types. These are the contact network (CN) of the user, shared contact with users outside of the network (FR), shared subscriptions (SBN), shared subscribers (SBR), and shared favorite videos (VID). Additional information about the data can be found in [6]. The basic edge statistics can be found in Table I.

B. Disease-Gene Network

The disease-gene (DG) network was constructed from three individual data sets. As the name suggests, this network has two distinct node types, diseases and genes, with four edge types connecting them. The diseases are classified by Disease Ontology (DO) codes and the gene names are based on the HUGO Gene Nomenclature. Genetic association (G) links exist between diseases and genes in a bipartite fashion and represent known disease-gene associations extracted from the Online Mendelian Inheritance in Man (OMIM) database, Swiss-Prot, and the Human Protein Reference Database (HPRD). Protein-protein interaction (PPI) links connect pairs of genes in accordance with combined physical interaction data collected from HPRD, the Online Predicted Human Interaction Database (OPHID), and studies by Rual [7] and Stelzl [8]. Further details about these datasets, which are publicly available, can be found in [9].

Additionally, disease pairs are connected by a phenotypic (P) link if they are significantly co-morbid in real patients. For our purposes, co-morbidity can be broadly defined as co-occurrence in the same patients significantly more than chance. We included edges between disease pairs for which the co-occurrence (joint probability) is significantly greater at 95% confidence than the random expectation based on population prevalence of the diseases (product of marginal probabilities), as determined by a two-proportion z-test. Disease co-morbidity was calculated from real patient medical histories collected from a group of 77 physicians within a regional health system. This includes data for the last 12 years, from 1997 to 2009, with a total of 5.5 million visits for approximately 700,000 patients. Each data record is a single visit represented by an anonymized patient ID and a primary diagnosis code, as defined by the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). For consistency with the first dataset, the ICD-9-CM codes have been converted to Disease Ontology codes based on mappings provided within the DO coding. Due to the hierarchical structure of the Disease Ontology (DO) codes, some disease pairs are connected by a "family" (F) link, where one disease is a more general hierarchical parent of the other. For example, *Toxic pneumonitis* is a type of *Pneumonia*. The family links supersede and replace phenotypic links, since these pairs are not separate diagnoses. For our experiments, family links form a separate fold which is always in the training set and are not predicted; they are simply an artifact of the code structure.

The disease-gene network consists of 703 diseases and 1,132 genes, and edge statistics can be found in Table I.

Youtube	
Nodes	
Users	15,088
Edges	
CN	76,765
FR	1,940,806
SBN	2,239,440
SBR	5,574,249
VID	3,797,635

Disease-Gene	
Nodes	
Diseases	703
Genes	1,132
Edges	
G	10,483
P	74,523
PPI	2,450
F	3,279

Climate	
Nodes	
Locations	1,701
Edges	
GH	249,322
VWS	71,458
PW	50,835
RH	25,375
SST	132,469
SLP	175,786
HWS	31,615

TABLE I
NODE AND EDGE COUNTS FOR THE YOUTUBE, DISEASE-GENE, AND CLIMATE NETWORKS.

C. Climate Network

The climate network used in the paper is constructed from 5° x 5° latitude-longitude gridded climate data, where each node is a physical location and edges represent similarity with respect to one of seven climate variables. The variables included are temperature (SST), sea level pressure (SLP), horizontal (HWS) and vertical (VWS) wind speed, precipitable water (PW), relative humidity (RH), and geopotential height (GH), each of which is represented as a distinct edge type. Similarity is measured in terms of Pearson correlation, with a threshold of 0.3. Every edge type can overlap with every other; a single node pair may have up to seven edges. Additional information about this network and the raw data can be found in [10].

Overall, the climate network included 1701 location nodes, and nearly all of which have some edges of all seven types. Edge details are provided in Table I.

D. Evaluation Framework

For all experiments, we use a 10-fold cross-validation stratified edge holdout scheme. We chose holdout evaluation since longitudinal data was either not available or not relevant for these networks. Link prediction is evaluated for each edge type x separately on all eligible node pairs (s, t), where eligibility is defined as follows:

1) An edge of type x is not already present between s and t in the training set.
2) It is possible for s and t to have an edge of type x, according to domain rules (i.e. two disease cannot have a protein-protein interaction in the disease-gene network).

Link prediction performance is evaluated separately for each edge type using area under the receiver operating curve (AUROC).

III. Unsupervised Methods

There are many existing approaches to link prediction for standard networks with homogeneous edges, formulated for various link formation hypotheses. A survey of these methods is provided in [1]. In Section III-A, we briefly describe the unsupervised link prediction methods used in our experiments, which serve as a performance baseline and also as features for the supervised approach. We then introduce a novel unsupervised method for heterogeneous link prediction in III-C.

A. Homogeneous Link Prediction

1) Neighborhood Methods: Many traditional link prediction scores are derived from the immediate node neighborhoods. The *preferential attachment* [11], [12] link prediction score for a node pair is the product of their degrees. *Common neighbors* [13] is another simple method which counts the common neighbors of s and t, which is the equivalent to the number of paths of length 2 between the nodes. *Jaccard's coefficient* [14] is the number of common neighbors divided by the total combined number of neighbors of both nodes. Another variation of common neighbors is the *Adamic/Adar* measure [15], which weights the impact of neighbor nodes inversely with respect to their total number of connections. Specifically,

$$\text{score}(s,t) = \sum_{n \in N_s \bigcap N_t} \frac{1}{log(|N_n|)} \tag{1}$$

where N_x is the set of common neighbors of node x. This inverse frequency approach is based on the assumption that rare relationships are more specific and have more impact on similarity.

2) Path Methods: A second class of link prediction methods are calculated based on paths between nodes. The *PageRank* algorithm of Google fame, first introduced in the academic sphere in [16], represents the significance of a node in a network based on the significance of other nodes that link to it. If we assume that linking to nodes that are important is desirable, an assumption implicit in preferential attachment prediction, then the PageRank of the target node represents a useful statistic. For our experiments, we perform the original, unoptimized PageRank calculation iteratively, checking for convergence of the vector of PageRank scores by calculating the Pearson correlation coefficient, r. After $r < 0.85$, we stop iterating and use the scores. Convergence generally requires under 10 iterations.

Rooted PageRank [1] is another link predictor derived from the original PageRank in which prediction outputs correspond to the probability of visiting the target node in the prediction during a random walk from the source. A parameter α, the probability of restarting the walk at the source, allows the walker to avoid getting trapped in directed networks or dense areas. We use $\alpha = 0.15$. Again, prediction scores are determined after the walks converge. Especially with low to moderate values of α, this may take many walk steps. In addition to the parameter, the rate of convergence depends on the size and local density of the network. In our implementation of RootedPageRank, we perform 100,000 steps at a time, checking each time whether or not $r < 0.85$.

The *PropFlow* predictor introduced in [5] is a path-based predictor that models the link prediction score as being propagated radially outward from the source. Starting from the source node with a score of 1, all neighboring nodes are given an equal share of the score (in the unweighted case), or $1/|N_s|$. The scores continue outward, summing together for nodes which are reached by multiple paths. For our experiments, we limit the path search to length 10.

B. Bipartite Link Prediction

Recall that in the disease-gene network, the genetic edges have a bipartite structure. While preferential attachment and the path methods apply naturally to bipartite networks, the common neighbor methods are triangle-based and require modification. Note that for nodes s and t, each common neighbor n belongs to a unique path (s, n, t) of length two from s to t. In a bipartite network, nodes are always connected by paths of odd length. Thus, to extend common neighbors link prediction to bipartite networks, we simply count the number of unique paths $(s, n1, n2, t)$ of length three from s to t. Similarly, we formulate bipartite Jaccard's coefficient as the number of unique paths of length three from s to t divided by the total number of unique paths of length three starting at either s or t. For Adamic/Adar, we replace the log term with

$$log(|N_{n1} \bigcup N_{n2}|)$$

which, in bipartite networks, is equivalent to

$$log(|N_{n1}| + |N_{n2}|)$$

C. Multi-Relational Link Prediction for Heterogeneous Networks

The link prediction methods described in III-A have no direct applicability to heterogeneous information networks other than treating all nodes and edges equally, which can be detrimental to their performance for many reasons. Different types contain different information by nature, and various combinations introduce different amounts of evidence to the link prediction task. This is particularly troublesome when node or link types have very different frequency or distribution, which is clearly the case in the real world networks we use. Even if these barriers could be overcome, considering all relationships equally provides no information about the type of link being predicted.

We now introduce a novel multi-relational link prediction (MRLP) method for heterogeneous information networks which addresses these shortcomings to predict the location and type of new edges. The most important component of the MRLP method is an appropriate weighting scheme for different edge type combinations. The weights are determined

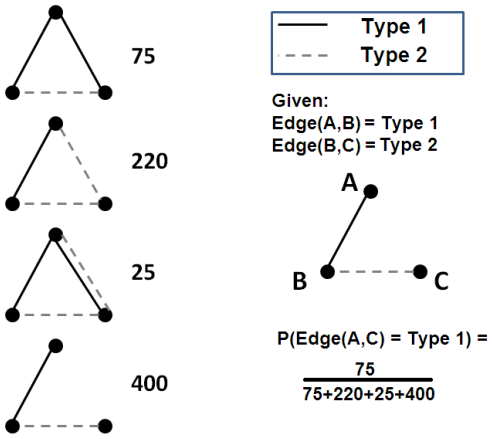


Fig. 1. This toy example demonstrates how to calculate the probability of a given edge type closing a partial triad structure based on triad census counts. These numbers do not represent a real network.

by counted the occurrence of each unique 3-node substructure in the network, traditionally called triad census [17] and more recently defined as counting 3-node graphlets [18]. The triad census trivially extends to heterogenous networks; the only difference is the number of unique structures. The triad census also provides the probability of each structure, which further translates to the probability that a partial triad is closed by each edge type. A pictorial example is shown in Figure 1. This conveniently translates to a non-arbitrary, data-justified weighting scheme. To account for frequency disparity, the probabilistic weights are normalized by the marginal probabilities of the edge types involved.

This multi-relational link prediction (MRLP) approach is a weighted extension of the neighborhood methods. Nodes s and t form a partial triad with each common neighbor $n \in N_s \bigcap N_t$, and each partial triad provides a probabilistic weight based on the triad census. We can simply add the weights, which is equivalent to weighted common neighbors. Prediction scores are calculated individually for each link type of interest. Formally, the prediction score for edge type x between nodes s and t is

$$\text{score}_x(s,t) = \sum_{n \in N_s \bigcap N_t} w_n \tag{2}$$

where

$$w_n = \frac{\sigma P(\text{edge\_type}(s,t) = x | \text{pattern}(s,n,t))}{P(\text{edge\_type}(s,n)) P(\text{edge\_type}(t,n))} \tag{3}$$

in which $\text{pattern}(s,n,t)$ describes the node and edge type pattern of the network path (s,n,t). Also,

$$\sigma = \begin{cases} 1 & P(\text{edge\_type}(s,t) = x | \text{pattern}(s,n,t)) > P(x) \\ 0 & P(\text{edge\_type}(s,t) = x | \text{pattern}(s,n,t)) = P(x) \\ -1 & P(\text{edge\_type}(s,t) = x | \text{pattern}(s,n,t)) < P(x) \end{cases} \tag{4}$$

where the sign is determined by statistical comparison rather than numerical. Statistical significance is determined by a two-tailed two proportion z-test with 99% confidence. As mentioned earlier, the denominator of the weight term is a normalization factor to account for the frequency disparity between edge types. The weighting scheme can suffer from the "zero frequency" or low frequency problem, which is particularly problematic in networks with a disproportionately large number of object and relationship types or many overlapping type combinations. Of course, the best solution is a larger sample, but this is often not available in practice. The problem, which is common to other probabilistic models such as Naive Bayes, can be combatted with many existing approaches such as smoothing operations. In our experiments, we set $\sigma = 0$ if $P(\text{edge\_type}(s,t) = x | \text{pattern}(s,n,t)) < 5$ or $P(x) < 10$, which corresponds to frequencies too low for a valid z-test. We assume that due to very low frequency, removing the influence of these patterns does not substantially effect the performance.

Equation 2 can be extended to include the inverse frequency principle of the Adamic/Adar measure, since it has been shown to increase performance in many cases. The integration is direct except that the degree is counted only with respect to the relevant node types. The prediction score becomes

$$\text{score}_x(s,t) = \sum_{n \in N_s \bigcap N_t} w_n \frac{1}{\log \begin{cases} |N_n(t1)| & t1 = t2 \\ |N_n(t1)| + |N_n(t2)| & t1 \neq t2 \end{cases}} \tag{5}$$

where $t1 = \text{edge\_type}(s,n)$, $t2 = \text{edge\_type}(t,n)$, and $|N_n(y)|$ is the number of edges of n with edge type y.

The weighting scheme used by MRLP obviously assumes that the dependence structure of new links will be similar to existing links, which we consider to be a common and reasonable assumption. However, it may be beneficial to determine the weights based only on "recently" formed links when time-series data is available, rather than characterizing all existing links. This is simply because newly formed links may not have the same topology as older links. Previous studies have detailed and advocated longitudinal evaluation schemes in systems where links form dynamically [5], [19], using longitudinal training, label, and testing intervals. This same interval scheme can be used to calculate the weights, counting only the partial triads formed by label edges and their common neighbors.

D. Results

The unsupervised link prediction methods were applied to single-dimensional (one edge type) projections of each network for each relationship type. In all cases except for disease-gene associations, the networks were also one-mode (one node type) networks and the standard link prediction methods were directly applicable. The disease-gene association network is bipartite, so modified versions of the neighborhood methods were applied (see Section III-B). The multi-relational link predictor (MRLP) was applied to the heterogeneous networks with no projections required, but each link type is evaluated separately for meaningful comparison. The complete evaluation framework is described in Section II-D.

	PA	PR	RPR	PF	JC	CN	AA	MRLP
YouTube								
CN	0.865	0.776	0.920	0.925	0.781	0.783	0.784	**0.945**
FR	0.934	0.835	0.953	0.962	**0.988**	0.984	0.986	0.971
SBN	0.944	0.844	0.922	0.938	**0.982**	0.980	0.981	0.969
SBR	0.946	0.851	0.964	0.973	0.987	0.988	**0.989**	0.973
VID	0.957	0.868	0.901	0.904	0.968	0.971	**0.973**	0.943
Disease								
G	0.903	0.786	0.933	0.951	0.957	0.951	0.956	**0.974**
P	**0.943**	0.813	0.808	0.762	0.771	0.909	0.911	0.938
PPI	0.827	0.723	**0.888**	**0.888**	0.786	0.788	0.789	0.808
Climate								
GH	0.783	0.684	0.953	0.939	0.989	0.985	0.986	**0.991**
VWS	0.802	0.730	0.861	0.893	**0.954**	0.935	0.942	0.948
PW	0.717	0.648	0.986	0.985	**0.995**	0.990	0.992	**0.995**
RH	0.681	0.608	0.991	0.991	**0.995**	0.992	0.993	0.993
SST	0.776	0.700	0.922	0.935	**0.975**	0.956	0.962	0.973
SLP	0.698	0.627	0.958	0.965	0.985	0.979	0.981	**0.988**
HWS	0.731	0.644	0.984	0.987	**0.995**	0.990	0.992	0.993

TABLE II
AUROC OF THE UNSUPERVISED LINK PREDICTION METHODS ON THE YOUTUBE, DISEASE-GENE, AND CLIMATE NETWORKS, COMPARED SEPARATELY FOR EACH LINK TYPE. THE METHODS APPLIED ARE PREFERENTIAL ATTACHMENT (PA), PAGERANK (PR), ROOTED PAGERANK (RPR), PROPFLOW (PF), JACCARD COEFFICIENT (JC), COMMON NEIGHBORS (CN), AND ADAMIC/ADAR (AA), AND THE MULTI-RELATIONAL LINK PREDICTOR (MRLP). THE BOLD NUMBERS INDICATE THE BEST PERFORMING UNSUPERVISED LINK PREDICTOR FOR EACH LINK TYPE. THE UNDERLINED NUMBERS INDICATE THE BETTER PERFORMANCE BETWEEN MRLP AND ADAMIC/ADAR, THE METHOD WHICH IT EXTENDS.

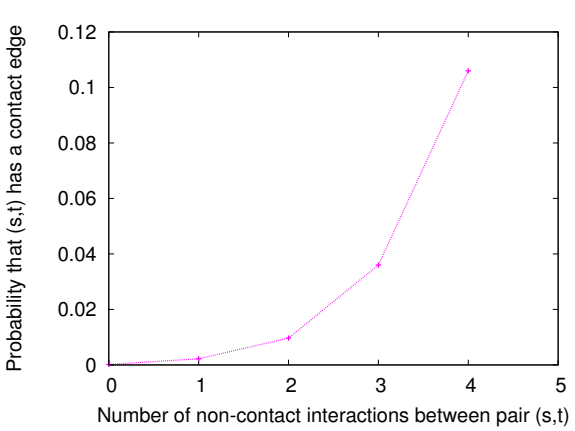
Performance results measured in AUROC for all of the unsupervised methods are shown in Table II. Methods in bold face indicate the best overall link predictor for the corresponding link type. First, we note that there is no universally dominant method, which is an expected result since unsupervised link prediction methods are domain-specific. Performance in all cases is fully dependent on how well the network of interest adheres to the pre-defined assumptions about link formation. For example, local neighborhood methods are clearly dominant in the YouTube network, a trait commonly observed in social networks [20], [21], [22], [23]. In the climate network, the Jaccard coefficient performs especially well, probably due to spatial autocorrelation [10]; that is, geographically proximate locations tend to have similar climate. In the disease network, each interaction type was best captured by a different method.

Like the other unsupervised methods, MRLP is also domain-dependent, bounded by the same principles as Adamic/Adar. In Table II, the underlined numbers indicate the better performance between MRLP and Adamic/Adar, the method which it extends. Comparison of these methods provides more meaningful insight into the benefit of capturing heterogeneous evidence. In the disease-gene and climate networks, MRLP performs comparable to or better than Adamic/Adar for all link types. This result strongly supports our basic assumption that diverse evidence can enhance link prediction, and further shows that our MRLP capture some of this potential improvement.

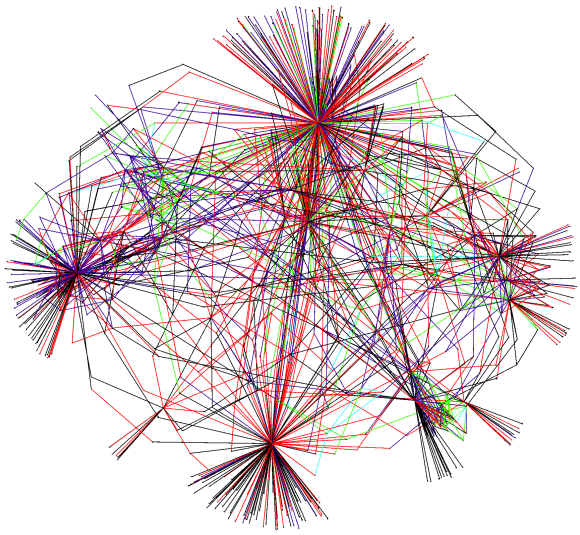
In the YouTube network, MRLP vastly improves performance for predicting the contact network (CN) between users, but degrades performance for the other relationships. We suspect that the huge benefit of additional evidence for the contact network is due to the relative scarcity of contact links, many of which are consumed by hubs. This is coupled with a high level of overlap with the other link types. In fact, 76% of all node pairs with a contact edge also have at least one other interaction type, and each additional overlapping interaction drastically increases the likelihood of a contact edges. A graph detailing the high probability overlap and an illustration on a sample of the contact network is provided in Figure 2. The sample is the full two-hop contact neighborhood of a randomly chosen seed user, including all connections between the neighbors reached by the crawl. Unfortunately, the redundancy we just observed may be harmful to the denser link types.The degraded performances on the other link types in the YouTube network suggest that MRLP does not deal well with noise introduced when additional link types do not provide beneficial information.

IV. SUPERVISED METHODS

A strong argument promoting a supervised approach to link prediction has been presented by previous work [5]. As we briefly mentioned in the previous section, the unsupervised link prediction methods can only perform well if the network link topology conforms to the *a priori* scoring function. This includes our multi-relational link predictor, which is limited by the same assumptions as the Adamic/Adar measure. We conjecture that other unsupervised methods could be extended to heterogeneous networks using similar probabilistic weighting schemes, but all *a priori* methods will have this same limitation. In heterogeneous information networks, the issue is exacerbated by the fact that multiple relationships may have vastly different formation mechanisms within the same system. For many real problems, such as the disease-gene network described in this paper, it is very unlikely that heterogeneous information types will share the same properties or distribution. In fact, even a single link type may form based on more than one mechanism. This supports the



(a) The probability of a contact edge for a given node pair (s, t) drastically increases when other interactions are present.



(b) This sample from the contact network illustrates the high prevalence of overlapping interactions when a contact relationship is present. Black edges indicate edges which are contact-only. Red, purple, green, and blue edges correspond to 1,2,3,or 4 additional interactions types, respectively.

Fig. 2. This figure demonstrates that the contact interaction in the YouTube network overlaps heavily with other interaction types, which may partially explain the strong performance enhancement resulting from heterogeneous evidence.

case for a supervised data-driven approach. A well-designed classification framework is not domain-specific, can support multiple decision boundaries, and can more flexibly combine the information provided by individual topological features. The link prediction problem is also extremely imbalanced, beyond the bounds of most problems studied by the imbalance community. While unsupervised methods ignore class distribution by definition, there are many well established approaches to combatting imbalance in a supervised setting.

The primary limitation of a supervised approach, compared to network methods, is the richness of the data representation. Choosing features that sufficiently represent the information inherent in the network topology is a serious challenge. Even fairly simple features describing the topology can increase exponentially with the number of node and edge types, as well as with the depth of the neighborhood being described. In this paper, we use a quite limited description of the heterogenous structure of the network, avoiding the exponential behavior. However, as our results will show, the benefits of the classification framework substantially outweigh the loss of information caused by the data transformation.

A. Classification

We study the performance of supervised classification in two contexts: first, the performance we can achieve using only homogeneous projections of the heterogenous information networks; second, the performance we can achieve when we incorporate information from different edge types into the classification scheme.

1) Feature Vector: For each projection of the network, the homogenous feature vector contains a combination of simple topological characteristics and standard unsupervised link prediction methods from [1] in the high performance link prediction (HPLP) approach provided in [5]. These include many of the unsupervised prediction methods already discussed earlier. In short, for both the source and target node, we include node degree and node PageRank with $d = 0.85$ [16]. For each pair, we also incorporate the common neighbors score, Jaccard coefficient, the Adamic/Adar method, the product of node degrees as a preferential attachment predictor, Rooted PageRank with $\alpha = 0.15$, PropFlow with $l = 10$, and the reciprocal of the shortest path from the source to the target up to 10 hops such that higher distances produce a 0 score.

2) Homogeneous Link Prediction: Within the 10-fold cross-validation evaluation paradigm, for each of the 10% testing evaluation, we first use 80% of the total edges (8 folds) to generate a network and construct the feature vectors for training. The remaining edges in the training fold, 10% of the total edges (the 9th fold), produce another network from which we generate the class labels. Specifically, every node pair (s, t) which does not have an edge in the 80% network is an instance and a feature vector is constructed based on the 80% network. If that node pair appears with an edge in the 10% label fold, it will be labeled as a positive instance (class 1). Otherwise, if no link exists between pair (s, t) in either division of the training fold, it will be labeled as a negative instance (class 0). The result is a standard data set feature vector with class labels. The testing set is produced in a similar manner. We then use the entire training fold to generate the feature vector for

testing, and the entire testing fold becomes the source for test labels. Just as it is unobserved in a standard classification task, the edges in the testing fold represent unobserved components of the total network topology with respect to the training procedure.

We can construct arbitrary models on this data; the only requirements we place on the classifier are that it accepts continuous features and binary class labels. We employ bagging [24] to reduce variance, but the chief challenge of the problem is class imbalance. To combat this without unfairly modifying the testing distribution, we undersample only the training set constructed within each fold so that the positives, links that actually exist, represent 25% of the data visible to the classifier. Each member of the ensemble for each fold sees all of the positives class instances from that fold and a different selection of negative class instances to achieve the indicated positive class representation. To increase both the speed and the performance of the classification step with respect to many alternative classification algorithms, we use random forests [25] within each bag. For each of the 10 bags, the forest contains 10 trees. We refer to this procedure as the high performance link prediction (HPLP) framework [5].

3) Heterogeneous Link Prediction: We now extend HPLP to the multi-relational case (MR-HPLP). To create the heterogeneous information classifier, we combine the data sets produced by each of the homogeneous projections. This method still loses information from the complete network, but allows us to employ readily available prediction scores. The combination of data sets is complicated by differences across the homogeneous projections. For the classification of each edge type, we must compose a separate amalgam of features for training from each of the other homogeneous projections. We can only test on unobserved edges in the projection in which we are predicting, but we can include in our feature vector whatever information we like from other projections. We can, for instance, include labels for edges that exist only in the test fold of other projections so long as they do not exist only in the test fold for the projection in which we are predicting. This requires generating all-pairs predictions for each projection separate from the testing process of the unsupervised and supervised homogeneous classification steps. Equation 6 provides an example of the amalgamated feature vector for projection 2 where $x(i)$ designates a feature in projection i with Greek letter subscripts showing different features for a given projection, $y(i)$ designates a label in projection i, and the class occupies the final column.

$$x(1)_\alpha, x(1)_\beta, x(1)_\gamma, y(1), x(2)_\alpha, x(2)_\beta, \\ x(2)_\gamma, x(3)_\alpha, x(3)_\beta, x(3)_\gamma, y(3), y(2) \quad (6)$$

B. Results

The supervised link prediction results using both individual (HPLP) and multiple (MR-HPLP) relationships are shown in Table III. For convenient comparison, we include the best unsupervised result from Table II. It is immediately clear that the supervised framework is dominant, outperforming the best unsupervised methods in all cases but one. Similarly, in all cases but one, MR-HPLP performs comparable or better to HPLP in all cases. The reason for the exceptions is not currently apparent, and is an avenue for future work. In many cases, the improvement gained by using MR-HPLP is somewhat minimal. This is not necessarily a detriment, and it is possible that the performance is nearing the saturation point for some interaction types. Also, on the other hand, MR-HPLP also shows low or negligible losses, which indicates better robustness to noise.

	Best (Unsupervised)	HPLP	MR-HPLP
YouTube			
CN	0.945	0.957	**0.973**
FR	0.988	**0.992**	**0.992**
SBN	0.982	**0.996**	**0.996**
SBR	0.989	**0.996**	**0.996**
VID	0.973	**0.985**	0.984
Disease			
G	0.974	**0.992**	0.988
P	0.943	**0.958**	**0.958**
PPI	0.888	0.898	**0.902**
Climate			
GH	0.991	0.986	**0.993**
VWS	0.954	0.966	**0.967**
PW	0.995	0.995	**0.996**
RH	0.995	0.996	**0.996**
SST	0.975	0.982	**0.983**
SLP	0.988	0.985	**0.992**
HWS	0.995	0.995	**0.996**

TABLE III
AUROC COMPARISON OF SUPERVISED LINK PREDICTION USING FEATURES CONSTRUCTED FROM EITHER THE HETEROGENEOUS (MR-HPLP) NETWORKS OR HOMOGENEOUS PROJECTIONS (HPLP) FOR EACH LINK TYPE. THE (BEST) UNSUPERVISED PERFORMANCE IS ALSO PROVIDED FOR EASY REFERENCE; THE SPECIFIC METHOD VARIES ACROSS TYPES. THE BOLD NUMBERS INDICATE THE OVERALL BEST LINK PREDICTOR FOR EACH LINK TYPE.

V. DISCUSSION

In this paper, we introduced two approaches to the link prediction problem in heterogeneous information networks. Our unsupervised multi-relational link predictor (MRLP) is an extension of the common Adamic/Adar approach. We also used heterogeneous topological features within a supervised framework for high performance link prediction (HPLP). Using experiments on three real-world networks from diverse domains, we discussed the performance trends and tradeoffs and conclude that supervised link prediction is the superior approach to link prediction in non-trivial networks, heterogeneous or not.

While our MRLP is only one of many ways to formulate or extend a topological link predictor for multi-relational data, we believe our experiments demonstrate some important trends of general interest. There is certainly information to be gained by integrating heterogeneous data, especially when homogeneous data is sparse. However, even when weighted, extended, and otherwise modified, unsupervised link predictors are still domain-specific and inflexible. This limitation is magnified in heterogeneous information networks. The most compelling natural network problems, such as human social

behavior or cell biology, are extraordinarily complex systems within systems. In these domains, it has already become unreasonable to try and describe such complexity with a single topological metric. Simple metrics remain very powerful, but now as features rather than models.

The results in this paper clearly support the shift toward supervised link prediction, which is consistently dominant in the homogeneous and heterogeneous networks. While the overall impact of heterogeneous features was modest in our supervised experiments, our integration of the heterogeneous features was also quite simplistic. Of course, the framework described is not limited to the features used. Developing supervised methods and features which efficiently capture the richness of heterogeneous information networks is perhaps the next logical step from the observations in this paper.

ACKNOWLEDGMENT

Research was sponsored in part by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 and in part by the National Science Foundation (NSF) Grant BCS-0826958. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *CIKM03: Proceedings of the twelfth international conference on Information and knowledge management*, 2003, pp. 556–559.

[2] H. Kautz, B. Selman, and M. Shah, "Referral Web: combining social networks and collaborative filtering," *Communications of the ACM*, vol. 40, no. 3, pp. 63–65, 1997.

[3] J. Han, "Mining heterogeneous information networks by exploring the power of links," in *Discovery Science*. Springer, 2009, pp. 13–30.

[4] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge Univ Pr, 1994.

[5] R. Lichtenwalter, J. Lussier, and N. Chawla, "New perspectives and methods in link prediction," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 243–252.

[6] L. Tang, X. Wang, and H. Liu, "Uncovering Groups via Heterogeneous Interaction Analysis," in *IEEE International Conference on Data Mining*, 2009, pp. 503–512.

[7] J. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. Berriz, F. Gibbons, M. Dreze, N. Ayivi-Guedehoussou *et al.*, "Towards a proteome-scale map of the human protein–protein interaction network," *Nature*, vol. 437, no. 7062, pp. 1173–1178, 2005.

[8] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen *et al.*, "A human protein-protein interaction network: a resource for annotating the proteome," *Cell*, vol. 122, no. 6, pp. 957–968, 2005.

[9] P. Radivojac, K. Peng, W. T. Clark, B. J. Peters, A. Mohan, S. M. Boyle, and S. D. Mooney, "An integrated approach to inferring gene-disease associations in humans," *Proteins*, pp. 1030–1037, 2008.

[10] A. Pelan, K. Steinhaeuser, N. V. Chawla, D. A. de Alwis Pitts, and A. R. Ganguly, "Empirical Comparison of Correlation Measures and Pruning Levels in Complex Networks Representingthe Global Climate System," in *IEEE Symposium Series on Computational Intelligence and Data Mining*, 2011.

[11] A. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A: Statistical Mechanics and its Applications*, vol. 311, no. 3-4, pp. 590–614, 2002.

[12] M. Newman, "Clustering and preferential attachment in growing networks," *Physical Review E*, vol. 64, no. 2, 2001.

[13] ——, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 2, p. 404, 2001.

[14] G. Salton and M. McGill, "Introduction to modern information retrieval," *New York*, 1983.

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

[16] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine* 1," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[17] P. Holland and S. Leinhardt, "Detecting structure in sociometric data," *American Journal of Sociology*, vol. 76, pp. 492–513, 1970.

[18] N. Pržulj, D. Corneil, and I. Jurisica, "Modeling interactome: scale-free or geometric?" *Bioinformatics*, vol. 20, no. 18, p. 3508, 2004.

[19] J. O'Madadhain, J. Hutchins, and P. Smyth, "Prediction and ranking algorithms for event-based network data," *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 2, pp. 23–30, 2005.

[20] N. Christakis and J. Fowler, "The spread of obesity in a large social network over 32 years," *New England Journal of Medicine*, vol. 357, no. 4, pp. 370–379, 2007.

[21] ——, "The collective dynamics of smoking in a large social network," *New England journal of medicine*, vol. 358, no. 21, pp. 2249–2258, 2008.

[22] N. Friedkin and E. Johnsen, "Social influence and opinions," *The Journal of Mathematical Sociology*, vol. 15, no. 3, pp. 193–206, 1990.

[23] M. McPherson, L. Smith-Lovin, and J. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, pp. 415–444, 2001.

[24] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[25] ——, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.