Soc. Netw. Anal. Min.
DOI 10.1007/s13278-012-0068-6

ORIGINAL ARTICLE

Supervised methods for multi-relational link prediction

Darcy Davis · Ryan Lichtenwalter · Nitesh V. Chawla

Received: 13 November 2011 / Revised: 21 March 2012 / Accepted: 4 April 2012

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 unsupervised link prediction. We develop supervised learning approaches for relationship (link) prediction in multi-relational networks, and demonstrate that a supervised approach to link prediction can enhance performance. 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.

1 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 (Liben-Nowell and Kleinberg 2003). 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 (Kautz et al. 1997). 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 (Barabási 2003; Scott 2011; Raeder and Chawla 2011; Han 2009; Kas et al. 2012). Structures of this type are broadly defined as heterogeneous information networks (Han 2009), an umbrella term, which encapsulates multi-mode, multi-dimensional, multi-relational, and bipartite networks (Wasserman and Faust 1994). Link prediction in these networks has typically been performed by treating all relationships equally or by studying each relationship separately 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

D. Davis · R. Lichtenwalter · N. V. Chawla (✉)
University of Notre Dame, 384 Fitzpatrick Hall,
Notre Dame, IN, USA
e-mail: nchawla@nd.edu

D. Davis
e-mail: ddavis4@nd.edu

R. Lichtenwalter
e-mail: rlichten@nd.edu

Published online: 29 April 2012

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.

1.1 Contributions

We describe and evaluate three new methods for the link prediction task in heterogeneous information networks. In Sect. 2, we describe three real-world heterogeneous data sources and our evaluation framework. In Sect. 3, 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 Sect. 3.3. 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 Sect. 4, we show that, just like homogeneous problems (Lichtenwalter et al. 2010), 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 Sect. 5.

2 Data

We use three real-world heterogeneous information networks to demonstrate the methods in this paper. The networks were chosen from different domains and are considerably divergent in structure and relationship types, which will support the generality of our conclusions.

2.1 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 (Tang et al. 2009). The basic edge statistics can be found in Table 1.

2.2 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 et al. (2005) and Stelzl et al. (2005). Further details about these datasets, which are publicly available, can be found in (Radivojac et al. 2008).

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. 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 data set, 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

Table 1 Network Details

YouTube		Disease-Gene		Climate	
Nodes		*Nodes*		*Nodes*	
Users	15,088	Diseases	703	Locations	1,701
		Genes	1,132		
Edges		*Edges*		*Edges*	
CN	76,765	G	10,483	GH	249,322
FR	1,940,806	P	74,523	VWS	71,458
SBN	2,239,440	PPI	2,450	PW	50,835
SBR	5,574,249	F	3,279	RH	25,375
VID	3,797,635			SST	132,469
				SLP	175,786
				HWS	31,615

Node and edge counts for the YouTube, disease-gene, and climate networks

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 1.

2.3 Climate network

The climate network used in the paper is constructed from $5^\circ \times 5^\circ$ 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 (Pelan et al. 2011).

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 1.

2.4 Evaluation framework

For all experiments, we use a 10-fold cross-validation stratified edge holdout scheme; that is, each fold contains a random 10% of the edges of each type, preserving the distribution. While longitudinal evaluation (i.e. using a training network observed at an earlier point of time than the testing network) is preferable for dynamic networks, time-series information was either not available or not relevant for the networks in this study. 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 characteristic curve (AUROC). We also evaluate area under the precision-recall curve (AUPR) (Raghavan et al. 1989). The separate evaluation of each edge type has some similarity to the one-versus-all framework for multi-class classification. Unlike traditional multi-class problems, edge types in our networks are not mutually exclusive, so there is no need to choose only one type for each node pair.

3 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 (Liben-Nowell and Kleinberg 2003). In Sect. 3.1, 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. In Sect. 3.2, we extend the unsupervised methods, which do not normally work in the bipartite case, so that they can be applied to the disease-gene associations. We then introduce a novel unsupervised method for heterogeneous link prediction in Sect. 3.3.

3.1 Homogeneous link prediction

The common topological link prediction methods can be divided into two types: neighborhood methods and path methods. The neighborhood methods are typically limited only to connections among the immediate neighbors of the source and target nodes, while path methods allow more global influences.

We now define some standard notation used to formalize the link prediction methods in this paper. The variables s and t refer to the source and target nodes, respectively. The neighbors of any node x, which consist of all other nodes connected to x, will be represented by the set N_x.

3.1.1 Neighborhood methods

Many traditional link prediction scores are derived from the immediate node neighborhoods. The *preferential*

attachment (PA) (Barabási et al. 2002; Newma 2001) link prediction score for a node pair is the product of their degrees. *Common neighbors* (CN) (Newman 2001) 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* (JC) (Salton and McGill 1983) 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* (AA) measure (Adamic and Adar 2003), which weights the impact of neighbor nodes inversely with respect to their total number of connections. Specifically,

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

This inverse frequency approach is based on the assumption that rare relationships are more specific and have more impact on similarity.

3.1.2 Path methods

A second class of link prediction methods are calculated based on paths between nodes. The *PageRank* (PR) algorithm of Google fame, first introduced in the academic sphere in (Brin and Page 1998), 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 ten iterations.

Rooted PageRank (RPR) (Liben-Nowell and Kleinberg 2003) 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* (PF) predictor introduced in (Lichtenwalter et al. 2010) 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.

3.2 Bipartite link prediction

Recall that in the disease-gene network, the disease-gene associations (type G) 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. One possibility to extend the common neighbor methods to the bipartite case, proposed by Huang et al. (2005), is to replace N_t with $\bigcap_{n \in N_t} N_n$. This formulation is not necessarily symmetric, i.e. in some cases $\text{score}(s,t) \neq \text{score}(t,s)$. We instead use symmetric formulations based on paths of length 3 between s and t. First, we note that for nodes s and t in a single-mode network, 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

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

3.3 Multi-relational link prediction for heterogeneous networks

The link prediction methods described in Sect. 3.1 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 by counting the occurrence of each unique 3-node substructure in the network, traditionally called triad census (Holland and Leinhardt 1970) and more recently defined as counting 3-node graphlets (Pržulj et al. 2004). 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. Specifically, for three nodes (s, n, t) and an edge type x, we first count all triads with the same pattern as (s, n, t), then count all triads with the same pattern with x added between s and t. We can determine $P(x \subset \text{edge\_type}(s,t)|\text{pattern}(s,n,t))$ by dividing the first count by the second count. This probability assumes that the observed pattern is correct except for the potential absence of type x, which simplifies the calculation. A pictorial example is shown in Fig. 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

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

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

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

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 the target pattern occurs less than five times or if type x occurs less than ten times, 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 of a neighboring node n only counts edges with the types as the edges connecting n to s or t. The prediction score becomes

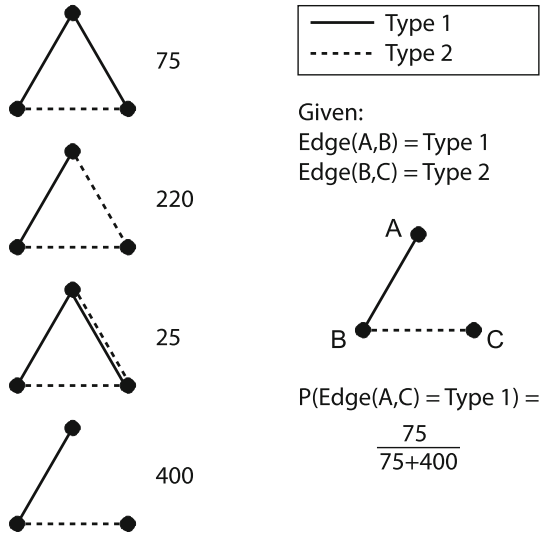


Fig. 1 Example: calculating the probabilistic weights for MRLP. This toy example demonstrates how to calculate the probability of a given edge type closing a partial triad structure based on triad census counts. The *left side* shows the subset of the triad census which conforms to the given constraints. On the *right* is the partial structure for which a probability is being calculated. When calculating probability, MRLP assumes that the observed structure is completely correct except for the potential absence of the prediction target. Thus, the *middle* two patterns do not affect the probability. Since no edge of Type 2 was observed between nodes A and C and Type 2 is not the target type, we assume that Type 2 is not present between A and C and the middle patterns are not relevant

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

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 defined in Eq. 3 remains unchanged. Unless otherwise noted, MRLP refers to the formulation in Eq. 5 in our experiments.

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 (Lichtenwalter et al. 2010; O'Madadhain et al. 2005), 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.

3.4 Results

We generated performance results for the homogeneous link prediction methods for both of the naive approaches: considering each edge type separately or treating edge types equally in a combined network. In the first case, which we will call the *homogeneous baseline*, we evaluated link prediction performance for each edge type on separate subnetworks which preserved all nodes but only preserved edges of one type. The standard link prediction methods were directly applicable, and prediction scores were generated individually for each edge type. The disease-gene association network is bipartite, so modified versions of the Jaccard coefficient, common neighbors, and Adamic/Adar methods were applied for those three values (see Sect. 3.2).

When treating all edge types as indistinguishable in a single network (the *combined baseline*), only a single set of prediction scores are produced. However, this set produces different performance values when evaluated with respect to each edge type, since the class values (link present or not present) may change. Since held out edges may fall in a position where another edge type is already present in the training network, we produce prediction scores for all pairs in the combined network, regardless of the edge status. For meaningful comparison, we maintain information about the edge types and evaluate each link type separately only on node pairs that did not have a link of that type in the training set. Fig. 2 demonstrates how separate performance values for each edge type can be produced by a single ranking.

We compared the homogeneous baseline methods to the multi-relational link predictor (MRLP) applied to the heterogeneous networks, again evaluating each edge type separately. The complete evaluation framework is described in Sect. 2.4.

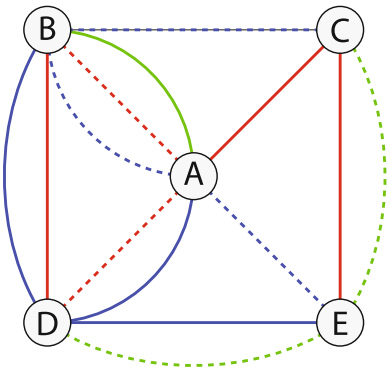
Performance results for the homogeneous baseline measured in AUROC and AUPR for all of the unsupervised methods are shown in Tables 2 and 3. Tables 4 and 5 similarly display results for the combined baseline. The performance for the MRLP method is provided in all tables for easy comparison. The values in bold face indicate the best overall prediction performance (per table) for the corresponding link type. From these results, we make two central observations:

1. No unsupervised method consistently performs well.
2. MRLP outperforms Adamic/Adar in most, but not all, cases.

An in-depth discussion of these points and other relevant trends is provided below.

First, we note the sometimes drastic difference in AUROC versus AUPR. While ROC and precision-recall curves are constructed from the same points, the area under the curves often differ widely, especially for imbalanced data (Davis and Goadrich 2006; Goadrich et al. 2004). While AUROC is the more widely used and understood method, it tends to skew toward high values in imbalanced data, and some researchers believe AUPR provides a more informative representation of the method's real performance and room for improvement (Davis and Goadrich 2006). We include both measures and observe that while the two methods sometimes disagree about the best method, both show the same basic trends, which we describe below.

It is immediately clear that the homogeneous baseline is a better approach than the combined baseline in the majority of cases. Values marked in italics in Tables 4 and 5 indicate the exceptions, where the performance of the combined baseline was higher than the corresponding value for the homogeneous baseline. The exceptions occur in the YouTube network only for the contact relationship, which is substantially improved throughout our results by any information from other edge types, distinguishable or not. A few minor exceptions are present for the climate network for preferential attachment and PageRank, both of which are weak models on the network and of little interest. The disease-gene network is more interesting case, with 12 cases where the homogeneous baseline performs better. While



Pair	Rank	Testing Class		
		■	■	■
(B,D)	1	-	-	0
(D,E)	2	0	-	1
(A,B)	3	1	1	-
(C,E)	4	-	0	1
(C,D)	5	0	0	0
(A,D)	6	1	-	0
(B,C)	7	0	1	0
(A,C)	8	-	0	0
(A,E)	9	0	1	0
(B,E)	10	0	0	0

AUROC(■)=0.70)
AUROC(■)=0.58
AUROC(■)=0.86

Fig. 2 Example: producing multiple AUROC values from a single ranking. The network on the *left* is a toy heterogenous network with 5 nodes and 3 edge types. *Solid lines* represent training edges and *dotted lines* indicate holdout edges for evaluation. This example assumes that all node pairs were ranked in descending order with respect to prediction scores produced by some algorithm. The ranking is fabricated and for demonstration purposes only. The testing classes are determined based on the network. *Dashes* indicate training edges, which are already known and do not impact performance. Class 1 contains the holdout edges, while class 0 comprises edges which are neither known nor held out. The AUROC can be calculated for each class column, as provided below the table. Each edge type produces a different value for the same ranking

Table 2 AUROC of the unsupervised methods (homogeneous baseline)

	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	**0.962**
P	**0.943**	0.813	0.808	0.762	0.771	0.909	0.911	0.867
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	**0.993**
VWS	0.802	0.730	0.861	0.893	**0.954**	0.935	0.942	0.949
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	0.992
SST	0.776	0.700	0.922	0.935	**0.975**	0.956	0.962	0.977
SLP	0.698	0.627	0.958	0.965	0.985	0.979	0.981	**0.990**
HWS	0.731	0.644	0.984	0.987	**0.995**	0.990	0.992	0.993

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

Table 3 AUPR of the unsupervised methods (homogeneous baseline)

	PA	PR	RPR	PF	JC	CN	AA	MRLP
YouTube								
CN	0.002	0.000	0.008	0.007	0.002	0.021	0.025	0.047
FR	0.056	0.008	0.042	0.037	0.394	0.397	0.413	0.145
SBN	0.176	0.024	0.060	0.070	0.625	0.521	0.541	0.431
SBR	0.135	0.012	0.117	0.121	0.497	0.346	0.377	0.359
VID	0.192	0.020	0.029	0.025	0.306	0.299	0.306	0.207
Disease								
G	0.036	0.013	0.134	**0.161**	0.001	0.001	0.001	0.060
P	**0.591**	0.195	0.191	0.134	0.287	0.561	0.558	0.238
PPI	0.008	0.001	0.011	0.009	0.007	0.040	0.044	**0.045**
Climate								
GH	0.617	0.034	0.228	0.169	0.806	0.756	0.766	**0.835**
VWS	0.045	0.014	0.047	0.052	**0.312**	0.241	0.255	0.214
PW	0.028	0.007	0.190	0.164	**0.674**	0.444	0.481	0.606
RH	0.052	0.003	0.156	0.146	**0.634**	0.311	0.362	0.439
SST	0.210	0.024	0.139	0.128	0.502	0.412	0.442	**0.514**
SLP	0.041	0.020	0.217	0.208	0.669	0.653	0.667	**0.716**
HWS	0.036	0.004	0.106	0.114	**0.573**	0.358	0.408	0.474

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

Table 4 AUROC of the unsupervised methods (combined baseline)

	PA	PR	RPR	PF	JC	CN	AA	MRLP
YouTube								
CN	0.733	0.681	0.829	0.855	*0.866*	*0.819*	*0.828*	**0.945**
FR	0.843	0.750	0.823	0.851	0.921	0.899	0.903	**0.971**
SBN	0.909	0.805	0.848	0.869	0.959	0.954	0.956	**0.969**
SBR	0.819	0.718	0.821	0.855	0.910	0.886	0.892	**0.973**
VID	0.895	0.800	0.823	0.835	0.922	0.926	0.927	**0.943**
Disease								
G	0.829	0.632	0.826	0.886	0.554	0.821	0.857	**0.962**
P	***0.944***	0.807	*0.812*	0.770	0.759	0.909	0.911	0.867
PPI	0.708	0.689	**0.846**	0.859	0.746	0.777	*0.801*	0.808
Climate								
GH	0.769	*0.705*	0.901	0.888	0.977	0.954	0.957	**0.993**
VWS	0.652	0.626	0.693	0.698	0.747	0.757	0.758	**0.949**
PW	0.596	0.548	0.874	0.883	0.911	0.856	0.863	**0.995**
RH	0.541	0.489	0.858	0.875	0.901	0.836	0.843	**0.992**
SST	0.710	0.652	0.821	0.822	0.893	0.882	0.885	**0.977**
SLP	0.628	0.589	0.870	0.881	0.923	0.887	0.893	**0.990**
HWS	0.515	0.493	0.829	0.855	0.876	0.816	0.824	**0.993**

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. Cases where the combined baseline outperformed the homogeneous baseline are marked in italics

most of the difference are modest, the AUPR of rooted PageRank and PropFlow show a huge jump from 0.1 to 0.4 for genetic associations. We suspect that the bipartite structure of the homogeneous genetic network was overly limiting for these methods. In fact, since the edge types in the disease-gene network are all domain-restricted and

Table 5 AUPR of the unsupervised methods (combined baseline)

	PA	PR	RPR	PF	JC	CN	AA	MRLP
YouTube								
CN	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.047
FR	0.016	0.004	0.008	0.009	0.028	0.022	0.023	0.145
SBN	0.089	0.020	0.009	0.011	0.166	0.168	0.171	0.431
SBR	0.014	0.005	0.024	0.026	0.029	0.015	0.016	0.359
VID	0.054	0.013	0.014	0.013	0.041	0.072	0.072	0.207
Disease								
G	0.012	0.011	*0.400*	***0.408***	0.001	*0.007*	*0.010*	0.060
P	**0.574**	0.171	*0.192*	*0.152*	0.243	0.560	0.557	0.238
PPI	0.005	0.001	*0.018*	*0.016*	*0.013*	0.005	0.025	**0.045**
Climate								
GH	0.318	*0.041*	0.123	0.106	0.757	0.676	0.683	**0.835**
VWS	0.012	0.009	0.010	0.010	0.016	0.016	0.016	**0.214**
PW	0.031	0.005	0.024	0.022	0.059	0.053	0.054	**0.606**
RH	0.004	0.002	0.011	0.011	0.040	0.011	0.012	**0.439**
SST	0.101	0.020	0.038	0.035	0.094	0.134	0.137	**0.514**
SLP	*0.045*	0.017	0.081	0.080	0.165	0.110	0.113	**0.716**
HWS	0.006	0.002	0.011	0.011	0.030	0.015	0.016	**0.474**

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. Cases where the combined baseline outperformed the homogeneous baseline are marked in italics

non-overlapping, we believe that there is less noise introduced by treating them equally; i.e. that the edge types are structurally distinguishable to some useful degree.

Overall, there is no universally dominant method. This is expected since unsupervised link prediction methods are known to be domain-specific.They are static, a priori link formation models rather than a trained models. 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 (Christakis and Fowler 2007; Christakis and Fowler 2008; Friedkin and Johnsen 1990; McPherson et al. 2001). In the climate network, the Jaccard coefficient performs especially well, probably due to spatial autocorrelation (Pelan et al. 2011); 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 Tables 2, 3, 4 and 5, 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 climate and disease-gene networks, MRLP performs comparable to or better than Adamic/Adar in the strong majority of cases. This result supports our basic assumption that the relationships between diverse information types can enhance link prediction.

Despite gains for some edge types, MRLP is not ideal in all cases. For the phenotypic edge type in the disease-gene network, Adamic/Adar performs equally well on the homogeneous and combined baselines, while MRLP performs worse. This indicates that additional edge types are neither harming nor helping predictions, but the weighting scheme of MRLP is detrimental. Essentially, the best predictor of phenotypic links is the number of common neighbors and the link patterns are irrelevant. MRLP does not adapt well to this scenario.

In the YouTube network, MRLP vastly improves performance for predicting the contact network (CN) between users, but usually degrades performance for the other relationships compared the the homogeneous baseline. 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 Fig. 3. 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.

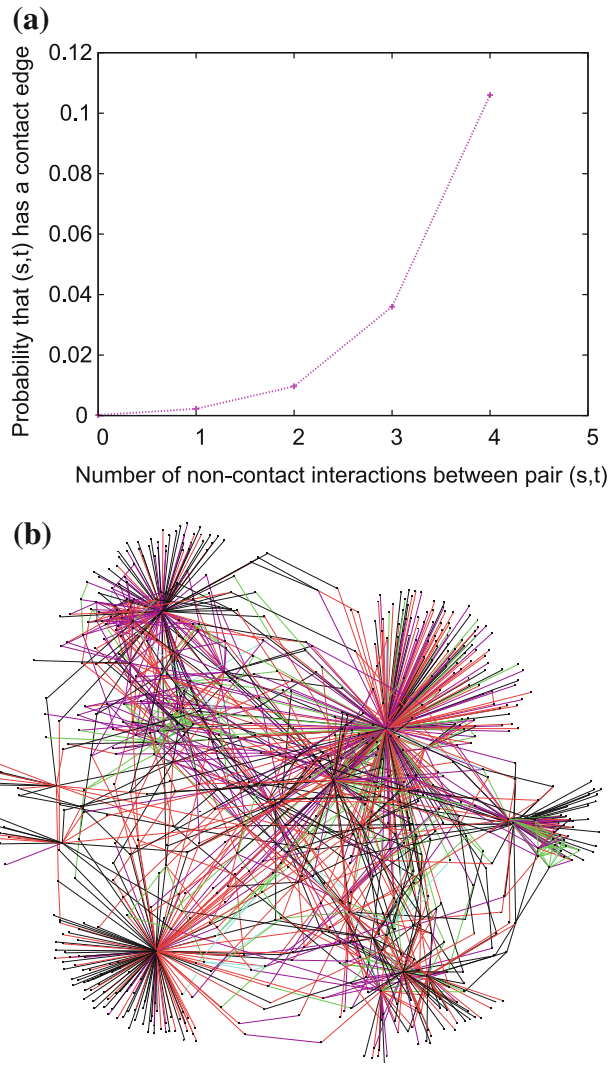


Fig. 3 This figure demonstrates that the contact interaction in the YouTube network overlaps heavily with other interaction types, which may partially explain the strong performance enhancement resulting from heterogeneous evidence. **a** The probability of a contact edge for a given node pair (s, t) 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 (color figure online)

4 Supervised methods

A strong argument promoting a supervised approach to link prediction has been presented by previous work (Lichtenwalter et al. 2010). 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. For example, counting common neighbors will only be successful in domains where sharing neighbors increases the chance of a relationship. While often effective in social networks, the common neighbors assumption may generalize poorly to domains with fewer small cycles, such as biological pathways or linguistic structures. This domain-specificity 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 static models 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, the heterogeneous information types have different properties and are not best predicted with any one metric. In fact, even a single link type may form based on more than one mechanism. This complexity supports the 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 reduced 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, such as the triad structures used by MRLP, can increase exponentially with the number of node and edge types, as well as with the depth of the neighborhood being described. The methods we will describe in this section use quite limited descriptions of the heterogeneous structure of the network, thus keeping the feature space manageable. However, as our results will show, the benefits of the classification framework can substantially outweigh the loss of information caused by the data transformation.

4.1 Classification weighted MRLP

One of the basic assumptions of the neighborhood based link predictors, including our MRLP method, is that evidence of a given type is linearly additive. That is, each additional common neighbor (or similar unit of evidence) adds a consistent amount to the score, regardless of the total number seen. Even Adamic/Adar, where nodes have different value, is linearly additive with respect to a given degree. In practice for homogeneous networks, it only matters that the real function of likelihood with respect to evidence be increasing, since most evaluation metrics are concerned with ranking rather than the distance between scores. Considering the success of neighborhood based methods, this has been a reasonable assumption for many networks of interest.

In the case of multiple edge types, this assumption often fails, especially if some types of evidence are stronger than others. An example from the climate network is shown in Fig. 4. Each series denotes the probability that a node pair (s, t) has a vertical wind speed (VWS) link with respect to the number of common neighbors connected to s and t by a given edge type. In this example, the distributions are not linear, are not always increasing, and vary widely between edge types. The predictive value of each additional piece of evidence (in this case, each common neighbor) is not constant nor even relatively static with respect to other edge types. This points to a clear weakness of the MRLP probabilistic weighting scheme, since it uses static weights.

To address these issues, we developed a *Classification Weighted version of MRLP* (CW-MRLP). First, we construct feature vectors from the triad structure counts employed by MRLP. We then train a classifier to distinguish between linked vs. non-linked node pairs based on the structures they are involved in. Like all experiments in this study, training is performed individually for each edge type. For each node pair (s, t) and each target edge type x, the corresponding feature vector is labeled class 1 if a link of type x exists between s and t, and class 0 otherwise. Information about edges between s and t are not included in the triad structures, since this information contains all class values. Instead, the class values of the non-target types are included as features. Figure 5 provides an pictorial example of how a network is converted into feature vectors and classes.

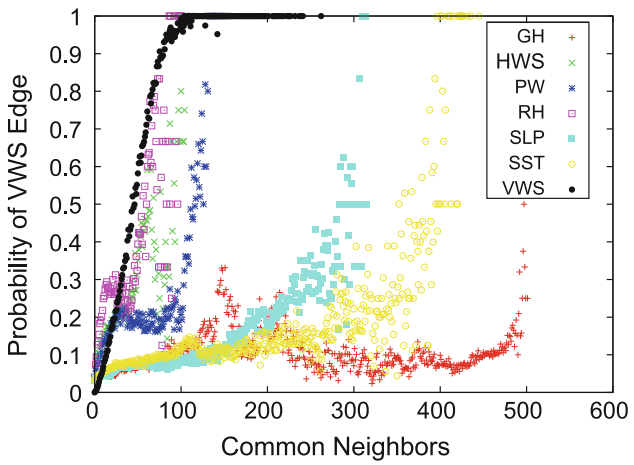


Fig. 4 Example: varying relative evidence between edge types. Each series denotes the probability that a node pair has a vertical wind speed (VWS) link with respect to the number of common neighbors connected by a given edge type

Consistent with our evaluation framework, edges are held out by flipping the label for the corresponding instance and edge type to class 0. A classification model is then learned on the data. The testing set is comprised of all feature vectors belonging to class 0 in the training set, but the feature vectors corresponding to held out edges were converted to class 1. Vectors of class 1 in the training set are not relevant since they correspond to previously known links. This setup is an interesting case of training on the testing set. However, there is no unfair advantage since all of the target instances are labeled with the wrong class. Also, this scenario accurately models the real-world problem, since most link prediction domains cannot distinguish between definite negatives versus link candidates.

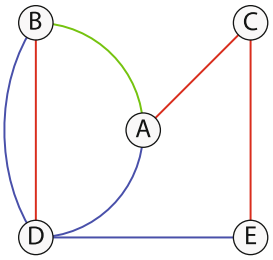
This supervised approach allows the classifier to determine the relative value of the features dynamically at each step while building the model, as well as allowing for multiple decision boundaries. CW-MRLP is still limited to the same type of features used by the neighborhood methods, but has the freedom to learning their value rather than making an a priori assumption.

4.2 Combining homogeneous predictors with HPLP

Another approach to supervised link prediction is to combine unsupervised link predictors scores as features in a classification scheme. We study the performance of this supervised approach in two contexts: first, the performance we can achieve using only the homogeneous baseline predictors; second, the performance we can achieve when we simultaneously incorporate scores from different edge types into the classification scheme.

4.2.1 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 (Liben-Nowell and Kleinberg 2003) in the *High Performance Link Prediction* (HPLP) approach provided in (Lichtenwalter et al. 2010). These include many of the unsupervised prediction methods already discussed



Pair	Pattern Counts							Training Class		
(B,D)	0	0	0	0	1	0	0	1	1	0
(D,E)	0	0	0	0	0	0	0	0	1	0
(A,B)	0	0	0	0	0	1	0	0	0	1
(C,E)	0	0	0	0	0	0	0	1	0	0
(C,D)	2	0	0	0	0	0	0	0	0	0
(A,D)	0	0	0	0	0	0	1	0	1	0
(B,C)	0	1	0	0	0	0	0	0	0	0
(A,C)	0	0	0	0	0	0	0	1	0	0
(A,E)	0	0	1	1	0	0	0	0	0	0
(B,E)	0	0	0	0	0	1	0	0	0	0

Fig. 5 Example: converting a heterogenous network into feature vectors and classes. Each instance is a unique node pair. The features represent possible edge patterns between the node pair and their common neighbors. While many additional patterns are possible with three edge types, these features are only included if they occur in the network. The classes describe the edge relationships present between source and target nodes. The class values of non-target edge types are used as additional features for training

earlier. In short, for both the source and target node, we include node degree and node PageRank with $d = 0.85$ (Brin and Page 1998). 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 ten hops such that higher distances produce a 0 score.

4.2.2 *Homogeneous link prediction*

Within the tenfold cross-validation evaluation paradigm, for each of the 10 % testing evaluation, we first use 80 % of the total edges (eightfolds) to generate a network and construct the feature vectors for training. The remaining edges in the training fold, 10% of the total edges (the ninthfold), 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). As with CW-MRLP, held out edges will be assigned to class 0. The result is a standard data set feature vector with class labels. The testing set is produced in a similar manner. We 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.

4.2.3 *Heterogeneous link prediction*

We now extend *HPLP to the Multi-Relational case* (MR-HPLP). The standard HPLP method accepts features that correspond to basic topological characteristics and unsupervised link predictors based on a collapsed view, or *projection*, of the heterogeneous network. This projection combines one or more relations from the original network into a network with only a single relation that simply indicates connectedness. In this particular case, we use HPLP to predict using a single relation at a time. MR-HPLP combines the data sets from each of the projections by concatenating the HPLP vectors so that the data set consists of a feature value for each HPLP feature for each projection. This method still loses information from the complete network since topological interactions between the underlying relations are lost. Its advantage is that it allows us to employ readily available prediction scores, because the features values are already computed as part of HPLP.

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

We provide an example of the MR-HPLP feature vector in Eq. 6. $x(i)$ designates a feature in projection i with Greek letter subscripts showing different features for a given projection, $y(i)$ designates the presence of an edge in projection i, and the class occupies the final column. Combining data sets in this manner is complicated by the fact that each projection has a different set of unobserved edges. We can only *test* on previously unobserved edges in the target projection for prediction, but it is fair to include in our feature vector whatever information we like from other projections. We can, for instance, include an extra feature indicating the presence of an edge of another relation type, because we may know of the presence of that relation without leaking information from testing data about the target relation. For the classification of each edge type, we must therefore compose a slightly different amalgam of features for training from each of the other homogeneous projections depending on what labels we can fairly use. This requires generating additional predictions for each projection separate from the testing process of the unsupervised and supervised homogeneous classification steps.

4.3 Classification framework

For all of the supervised methods that we describe in this section, 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. For our experiments, we employ bagging (Breiman 1996) 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 (Breiman 2001) within each bag. For each of the 10 bags, the forest contains ten trees. This framework has previously been shown to perform very well for link prediction in homogeneous networks (Lichtenwalter et al. 2010).

4.4 Results

The supervised link prediction results for classification weighted MRLP (CW-MRLP), combined individual link predictors constructed on homogeneous networks (HPLP), and combined predictors of all types (MR-HPLP) are shown in Table 6. For convenient comparison, we include the best unsupervised result from Tables 2 and 4. The supervised framework is more consistent in its performance, generally achieving higher scores than unsupervised methods. Also, classification weighted MRLP outperforms our original probabilistic formulation of MRLP for all types except protein-protein interactions. Due to the edge pattern restrictions in the disease-gene network, the feature vector for the PPI edges only contains two patterns, which is not ideal for the classification framework.

In most cases, we observed that a heterogeneous supervised approach (CW-MRLP or MR-HPLP) show the best or comparable-to-best performance overall. This does not hold for genetic associations; the reason is not currently apparent and is an avenue for future work. These approaches also have robustness to noise. Especially in the YouTube network, the improvement gained by using MR-HPLP versus homogeneous HPLP is often minimal. Aside from the contact network, which has been a consistent exception, it seems likely that heterogeneous information has little value in this network and is mostly noise. However, MR-HPLP handles the situation with little or no detriment to performance. Also, the losses suffered by MRLP in the YouTube network are not observed for CW-MRLP. The supervised version generally

Table 6 Performance of the supervised methods

	Best (Unsupervised)	CW-MRLP	HPLP	MR-HPLP
YouTube				
CN	0.945	0.969	0.957	**0.973**
FR	0.988	0.990	**0.992**	**0.992**
SBN	0.982	0.993	**0.996**	**0.996**
SBR	0.989	0.991	**0.996**	**0.996**
VID	0.973	0.979	**0.985**	**0.984**
Disease				
G	0.962	0.971	**0.992**	0.988
P	0.943	0.933	**0.958**	**0.958**
PPI	0.888	0.807	0.898	**0.902**
Climate				
GH	0.993	**0.997**	0.986	0.993
VWS	0.954	0.954	0.966	**0.967**
PW	0.995	**0.997**	0.995	0.996
RH	0.995	**0.996**	**0.996**	**0.996**
SST	0.975	**0.989**	0.982	0.983
SLP	0.990	**0.996**	0.985	0.992
HWS	0.995	**0.996**	0.995	**0.996**

The Best of 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

performs better than homogeneous Adamic/Adar values despite potential added noise.

5 Discussion

In this paper, we introduced three 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 combined a heterogeneous collection of traditional link predictors 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. Furthermore, it may be difficult to determine in advance as to which information types are useful, so robustness to noise is essential. The solution is supervised methods; 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 consistent in both homogeneous and heterogeneous networks. Our approach intentionally includes some incorrect labels in the training set, since real world link prediction domains rarely distinguish between known negatives and link candidates. Conventional wisdom suggests that supervised classifiers may overfit the flawed data and produce poor or unreliable results compared to unsupervised methods using correct-but-incomplete data. However, the perceived higher reliability of unsupervised methods doesn't hold up in practice against an appropriate supervised method. In reality, many important problems are subject to flawed data, and classification methods have adapted to the challenge. Our supervised approaches do not overfit and instead perform better within a robust validation framework.

While the support for supervised approaches was consistent, our experiments show that feature/model selection still isn't free. The neighborhood patterns employed by CW-MRLP were dominant in the climate network, while the other networks benefitted more from the diverse topological measures of HPLP. Furthermore, 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.

Acknowledgments 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

Adamic L, Adar E (2003) Friends and neighbors on the web. Soc Netw 25(3):211–230

Barabási A (2003) Linked: how everything is connected to everything else and what it means. Penguin Group, New York

Barabási A, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. Phys A Stat Mech Appl 311(3-4):590–614

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30(1–7):107–117

Christakis N, Fowler J (2007) The spread of obesity in a large social network over 32 years. New Engl J Med 357(4):370–379

Christakis N, Fowler J (2008) The collective dynamics of smoking in a large social network. New Engl J Med 358(21):2249–2258

Davis J, Goadrich M (2006) The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd international conference on machine learning ACM, pp 233–240

Friedkin N, Johnsen E (1990) Social influence and opinions. J Math Sociol 15(3):193–206

Goadrich M, Oliphant L, Shavlik J (2004) Learning ensembles of first-order clauses for recall-precision curves: a case study in biomedical information extraction. In: Inductive logic programming, pp 421–456

Han J (2009) Mining heterogeneous information networks by exploring the power of links. In: Discovery science, Springer, pp 13–30

Holland P, Leinhardt S (1970) A method for detecting structure in sociometric data. Am J Sociol 76(3):492–513

Huang Z, Li X, Chen H (2005) Link prediction approach to collaborative filtering. In: Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries, ACM, pp 141–142

Kas M, Carley K, Carley L (2012) Trends in science networks: understanding structures and statistics of scientific networks. In: Social network analysis and mining, pp 1–19

Kautz H, Selman B, Shah M (1997) Referral Web: combining social networks and collaborative filtering. Commun ACM 40(3):63–65

Liben-Nowell D, Kleinberg J (2003) The link-prediction problem for social networks. In: Proceedings of the 12th international conference on information and knowledge management, pp 556–559

Lichtenwalter R, Lussier J, Chawla N (2010) New perspectives and methods in link prediction. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 243–252

McPherson M, Smith-Lovin L, Cook J (2001) Birds of a feather: Homophily in social networks. Ann Rev Sociol 27:415–444

Newman M (2001) Clustering and preferential attachment in growing networks. Phys Rev E 64(2):025102

Newman M (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci 98(2):404

O'Madadhain J, Hutchins J, Smyth P (2005) Prediction and ranking algorithms for event-based network data. ACM SIGKDD Explor Newslett 7(2):23–30

Pelan A, Steinhaeuser K, Chawla NV, de Alwis Pitts DA, Ganguly AR (2011) Empirical comparison of correlation measures and pruning levels in complex networks representing the global climate system. In: IEEE symposium series on computational intelligence and data mining

Pržulj N, Corneil D, Jurisica I (2004) Modeling interactome: scale-free or geometric? Bioinformatics 20(18):3508

Radivojac P, Peng K, Clark W, Peters B, Mohan A, Boyle S, Mooney S (2008) An integrated approach to inferring gene–disease associations in humans. Proteins Struct Funct Bioinform 72(3): 1030–1037

Raeder T, Chawla N (2011) Market basket analysis with networks. In: Social network analysis and mining, pp 1–17

Raghavan V, Bollmann P, Jung G (1989) A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans Inform Syst (TOIS) 7(3):205–229

Rual J, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz G, Gibbons F, Dreze M, Ayivi-Guedehoussou N et al (2005) Towards a proteome-scale map of the human protein–protein interaction network. Nature 437(7062):1173–1178

Salton G, McGill M (1983) Introduction to modern information retrieval. McGraw-Hill, New York

Scott J (2011) Social network analysis: developments, advances, and prospects. In: Social network analysis and mining, pp 1–6

Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck F, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S et al (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122(6):957–968

Tang L, Wang X, Liu H (2009) Uncovering groups via heterogeneous interaction analysis. In: Proceedings of the 9th IEEE international conference on data mining, pp 503–512

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge university press, Cambridge