2010 IEEE International Conference on Data Mining Workshops

Link Prediction Across Multiple Social Networks

Muhammad Aurangzeb Ahmad, Zoheb Borbora, Jaideep Srivastava
Department of Computer Science
University of Minnesota
Minneapolis, USA
mahmad, zborbor, srivasta@cs.umn.edu

Noshir Contractor
Shool of Communication
Northwestern University
Chicago, USA
nosh@northwestern.edu

***Abstract*—The problem of link prediction has been studied extensively in literature. There are various versions of the link prediction problem *e.g.,* link existence problem, link removal problem, predicting edge weights over time etc. In this paper we describe a new type of link prediction problem called the Inter-network link-prediction problem where the task is to predict links *across* different networks. Thus given a set of nodes which participate in multiple networks the task is to determine if one can predict the edges that occur in one network by only using node attribute and edge information from other networks. We use insights from theories of evolution of social communication networks and the MTML framework to derive models which can be used to make link predictions across networks. For the experiments data from different *types* of social networks from a Massively Multiplayer Online Role Playing Game (MMORPG) is used.**

***Keywords*-Inter-Network link prediction, social networks, multiple social networks, link prediction**

I. Introduction

Humans are social creatures and interact with one another in a variety of manners such that human social networks are ubiquitous in nature. Such social networks can range from offline networks based on friendship or kinship ties to online networks in social networking websites like Facebook, LinkedIn, MySpace etc. Social Networks are most often represented as graphs or hypergraphs and there is an extensive body of literature on social networks[22]. Various predictive problems have been proposed for social networks, the link prediction problem is the problem of predicting links in a network which may form in the future between the nodes in the network. The link prediction problem was first proposed by Liben-Nowell and Kleinberg [13] and has been studied extensively since then[25]. The link prediction actually consists of a family or subproblem *e.g.,* predicting the existence of the link, type of the link, the strength of the link etc.

While the link prediction problem has been applied in many domains like social networks, protein-protein interaction, record linkage problem, web-link prediction etc, we restrict ourselves to application of link prediction in social networks although our technique can be applied in other domains as well. The link prediction problem has a wide range of applications in addition to the well known example of *e.g.,* recommendation systems [9][26], making recommendation to create mutually beneficial professional links [25], improve navigational efficiency of websites[25] etc.

For experiments and validation, data from an Massively Multiplayer Online Role Playing Game (MMO) EverQuest II (EQ2) is used for multiple *types* of networks were extracted from the game for the link prediction tasks. These networks form because of different types of social processes and represent networks of different nature. Most of the previous work on link prediction has used data from citation or collaboration [13] networks for link prediction. The link prediction problem is well studied in such types of datasets. Here we concentrate on networks which form as a result of different social processes. Previous studies of socializing [24] in virtual worlds suggest that causality in virtual worlds is similar to that in the real world. Consequently results in insights from studying virtual world may be applied to the offline world in some contexts.

The problem of link prediction actually consists of a family of prediction problems. To the best of our knowledge the previous literature on link prediction is restricted to link prediction within the same network. In this paper we propose a new problem, inter-network link prediction (INLP), which is the problem of predicting the formation of links across networks *i.e.,* given networks G_1 and G_2 the task is to use information from G_1 to make predictions about G_2 and vice versa. Link prediction techniques exploit various techniques like the attributes of the nodes, topological features of the graph or aggregate features of the nodes to make predictions about the links. The performance of some of these techniques can be enhanced by adding domain knowledge to these techniques. An oft neglected source of domain knowledge is social science theories which link back to network processes that may be going on in various social networks. In this paper we seek to employ insights from theories of social communication to enhance the inter-network link prediction task. Many of these theories propose the existence of 'Structural Signatures' (expected subgraphs) which are likely to be present in certain types of networks. To the best of our knowledge social science theories have not been applied before in improving link prediction. The

978-0-7695-4257-7/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDMW.2010.79

closest example that is work by L and Zhou [14] who used weighted versions of many well known local topological measures to make predictions and discovered that weighted versions perform worse than their unweighted counterparts. The implication being that the weak ties in the network play a significant role in link prediction.

There are multiple theories about how social communication networks evolve over time. Monge and Contractor [15] developed the Multi-Theoretical Multi-Level (MTML) Framework which synthesized insights from various theories of social communication and also identified a set of structural signatures associated with each type of theory. Insights from MTML are used to propose a model for predicting links across networks. The link prediction family of problems can be addressed through different frameworks, in this however paper we focus on the predictive power of topological features and use a machine learning approach similar to that of Hasan et al [4]. The contributions of this paper can be summarized as follows:

- Define a new link prediction problem, the inter-network link prediction problem, and propose a solution to the problem.
- Use of insights from social science theories to augment the process of link prediction and improve the results of link prediction.
- Define and address the problem of inter-network link prediction where network information from one network can be used to make link predictions in another network.

The rest of the paper is organized as follows: In section II we describe related work in the domain of link prediction and background from theories of social communication networks, in section IV we describe our proposed approach and in section III we describe the inter-network link prediction task. The dataset, experiments and results are described in section V and the conclusion is in section VI.

II. Related Work

A. The Family of Link Prediction Problems

The link prediction problem was first proposed by Liben-Nowell and Kleinberg [13] who used various graph proximity measures to make predictions about co-authorship networks in Physics. Rattigan et al [19] defined the problem of anomalous link discovery where the task is to discover links which may be "surprising" as compared to other links in the network. The motivation being that since the number of dyads that have to be evaluated for link prediction grows combinatorially as the network size grows it is more useful to concentrate on the surprising links. The link prediction problem consists of a family of prediction problems. While most classifications of the link prediction problem sub-divide it into two or three sub-problems [25], we give a more comprehensive classification of the problem as follows:

- Link Formation Prediction. (Does a link exist?) [13][4]
- Link Disappearance Prediction. (Will a current link disappear?) [20]
- Link Classification. (What is the nature of the link?)[25]
- Anomalous Link Discovery. (What are the unexpected links?) [19]
- Link Weight Prediction (Predict the change in the weight of link) [25]
- Time Series Link Prediction (Prediction which links will reoccur over time)[21]
- Link Regression. (How does a user rate an item?)[9]

A number of topology based measures have been used for the link prediction tasks, these include Newman′s common neighbors [17], Jaccard's Index, Adamic/Adar metric[3] etc. Murata and Moriyasu [16] extend these metrics for weighted graphs and use for link prediction in a Q&A system. Huang [10] proposed a graph topology based method which generalizes the clustering coefficient and defines the problem of link prediction as that of cycle completion in graphs. It should also be noted that the topology based formulation of the problem can also be described as the problem of matrix completion which can be accomplished by matrix factorization [12]. Topology based temporal metrics were employed by Potgieter et al [18] to increase the performance of link prediction techniques.

Clauset et al [6] describe a maximum likelihood based approach combined with Monte-Carlo Algorithm and fit a hierarchical model on a network graph to make link predictions. Kunegis and Lommatzsch [11] describe a framework for link prediction based on transformation of a graphs algebraic spectrum. Their approach generalizes many previous graph kernel based approaches for link prediction. Sharan et al [20] describe a method based on graph summary which can predict both link formation and disappearance. For a more detailed review of link prediction literature we refer the reader to a survey on link prediction by Xiang [25].

B. Theories of Social Communication Networks

In this section we adopt the word communication as used in the context of social networks and is defined in terms of flow of ideas, commodities, influence etc in a social networks [15]. There are many theories which describe how social communication networks evolve over time and how links form in these networks. The salient features of the most prominent theories are given below. These theories are based on hundreds of empirical studies in social science and have a solid empirical and theoretical grounding. We refer the reader to the text by Monge and Contractor [15] for a more detailed description of these theories.

These theories describe a different aspect of communication networks. In some social network some of the theories are more applicable than others *e.g.,* theories of balance may explain friendship networks better than say theories of contagion. Monge and Contractor [15] developed

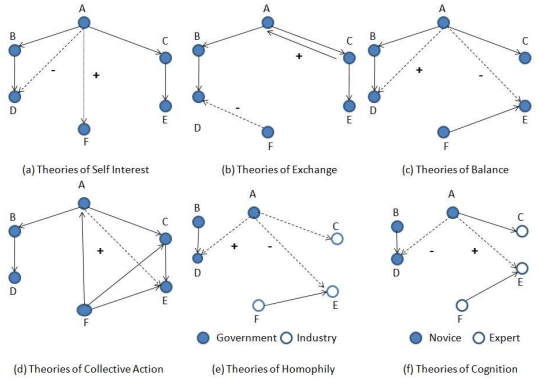

Figure 1. Structural Signatures of MTML

the Multi-Theoretical Multi-Level Framework which synthesized insights from the theories described above. There are certain archetypical behaviors which are found in many human networks which are expected to occur in networks where one type of social theory is at play versus another type of theory. These behaviors are: Exploring, Exploiting, Mobilizing, Bonding and Swarming.

The corresponding social theories for these are given in Table I. Based on these behaviors and theories they also identified network substructures or subgraphs that are likely to be associated with each type of theory. The corresponding structural signatures for these theories is given in Figure 1. The MTML framework has been adopted to determine the applicability of each type of these theories in various networks [7]. The models which are most commonly used for this purpose are the Exponential Random Graph Models (ERGM) or the p* family of models [22]. The main idea behind the ERGM model is that given a network and an expected set of structures (subgraphs) we determine that in the space of all possible graphs with the same number of nodes how likely is the observed network given the distribution of the expected substructures over all possible such graphs.

III. Inter-Network Link Prediction

In this section we describe inter-network link prediction problem (INLP). **Definition:** Given two graphs G_A and G_B with set of nodes $n_A \in N_A$ and $n_B \in N_B$, if $N_C = N_A \cup N_B$ and $N_A \cap N_B \neq \phi$., if $e_{Aij} \in E_A$ is the set of edges observed in the graph G_A then the task of inter-network link prediction is to predict $e_{Bij} \in E_B$ by using only the node $n_A \in N_A$, $e_{Aij} \in E_A$ or attribute $v(N_A)$ information from graph G_A.

As a practical example consider a case where network data is available from a social network amongst people who play golf together and additional information like demographics, profession, frequency of interaction etc is also available. There is also membership information available from another dataset for a subset of the people regarding trade but the edges in the trade network are not available. The task in this case would be to use the golf network and the additional available information to make predictions about the edges in the trade network. This type of information can also have practical applications like using it for marketing purposes etc.

IV. A Structural Signatures Based Approach

We now describe a structural signatures based approach for the link prediction problem. The main idea is to identify a set of substructures or subgraphs which are likely to be present in certain types of graphs which are known to be generated by certain social processes. We propose an algorithm, MTML Inter NeTwork Predictor (MINTP), for predicting link across networks. Before describing the algorithm in detail we first describe some background and motivations for this approach.

The MTML theory predicts the existence of certain substructures in networks which are driven by certain social processes. For a more detail exposition of this idea and the MTML theory we refer the reader to [15]. It should also be noted that the existence of these sub-structures are not independent from one another e.g., if we are considering only one type of network then certain types of structures are likely to occur in these networks.

On the other hand if we use information from multiple interacting networks then we would end up with different structures. The MTML theory also implies he transformation of certain types of sub-structures to other types. This information can be used for predictive purposes e.g., it could be the case that these transformations correspond to presence or absence of link formation that we are likely to see in the network. The following example can be used to illustrate this.

Consider Figure 2 which shows the evolution of graph G_i at time t_1, t_2, t_3, t_4. The subgraphs g_1 through g_n are the various subgraphs that are observed in the graph G_i. The notation $g_i \longrightarrow g_j$ denotes that graph g_i is gets transformed into graph g_j. It is clear from the figure that certain types of subgraph are being transformed into other types e.g., $g_1 \longrightarrow g_5$ in (t_1, t_2), $g_2 \longrightarrow g_7$ in (t_1, t_2) and (t_3, t_4), $g_3 \longrightarrow g_8$ in $(t_1, t_2), (t_2, t_3)$ and (t_3, t_4). These can also be automatically discovered by sequential pattern mining but since we are assuming that information about one network is not available we cannot apply sequential pattern mining. As a substitute however we are interested in the subgraphs which are predicted by the various social science theories and the transformation of these subgraphs into other subgraphs. The problem of finding such transformation can be represented as follows:

$$G_{t,t_\Delta+t} = \{g_i \longrightarrow g_j \| g_i \in G_t \wedge g_j \in G_{t_\Delta+t}\} \quad (1)$$

The task of making predictions can thus be defined in terms of determining how many such transformations $g_i \longrightarrow g_j$

Table I
THE MTML FRAMEWORK: SOCIAL DRIVERS FOR CREATING AND SUSTAINING COMMUNITIES

	Exploring	Exploiting	Mobilizing	Bonding	Swarming
Theories of Self Interest	+		-		
Theories of Cognition		+	+		+
Theories of Balance	-		+	+	
Theories of Exchange		+		+	
Theories of Contagion	+		+		
Theories of Homophily	-			+	
Theories of Proximity	-			+	+

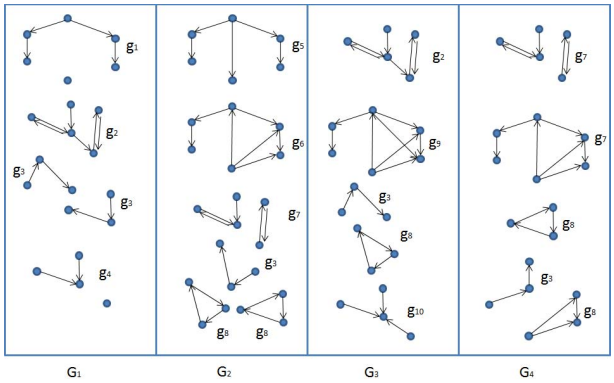

Figure 2. Examples of subgraph transformations

will result in transformation into cases where the link is formed between two nodes. To illustrate this again consider graphs g_6 and g_9 in Figure 2. From t_2 to t_3 an additional link is formed in the graph when $g_6 \longrightarrow g_9$. If we see the same transformation occurring in sufficiently large number of cases, as compared to such a transformation occurring in purely random graphs, then we can predict that this link is likely to form. We note that an alternative method would be to use graph generators to generate the underlying networks for comparison.

The outline of the algorithm is given below. The main idea is as follows: Given a graph G_A which contains the edge information and the node attribute information and another graph G_B which only contains the node information but no edge information, take the union of the graphs. It should be noted that the union does not include any edge information from G_B. Rewire the edges until the stopping criteria of adjacency correlation, details are given in the experiments section, is met. For the next k iterations perform the following procedure. Save the nodes in the graph in a lexicographical order and the randomize the ordering. Create or delete edges based on their likelihood of presence or absence according to their likelihood as described by MTML (see section **??**). The final predictions are based on taking the average of the network by running this procedure z times.

We note that in this algorithm k and z are the free parameters which describes the number of times these graphs have to be generated. For our experiments we used $k, z = 10,000$. If **T** represents the set of theories which are applicable to a particular domain then the MINTP Algorithm can be described in the figure below. The notation $e_A ij = (n_i \rightarrow n_j)$ implies that the edge $e_A ij$ exists between nodes n_i and n_j. in Graph A.

MINTP Algorithm:
Given: Graph G_A with edgeset E_A and nodeset N_A, Graph G_B with nodeset N_A
Task: Predict edges $e_{Bi} \in E_B$, $G_A \overset{E_A}{\rightarrow} E_B$.
$T \Rightarrow S = \{g_1, g_2, .., g_n\}$, (Substructures associated with **T**)
begin $G_P := (N_P = N_A \cup N_B, E_P = E_A)$
  while $A(G_P, G_B) \leq A(G_P, G_B)$ do
    Remove:$e_{Pij} := rand(E_P)$,
    $e_{Pij} = (n_i \rightarrow n_j), \exists e_{Aij} = (n_i \rightarrow n_j) \in E_A$
    Add:
    $e_{Pij} = (n_i \rightarrow n_j), n_i \in N_A, n_j \in N_B$ od
  for $i := 1$ to k do
    $n_i = rand(N_P), n_j = rand(N_P)$,
    P $= p(e_{Pij} \| c_1, c_2, ..., c_n)$
    if $\exists e_{Pij}$
      if **P** $< \tau$do
        $Remove : e_{Pij}$ od
    else
      if **P** $< \tau$
        $Add : e_{Pij}$ od
      od
    od
  end

One of the most important elements of the MINTP algorithm is to determine the conditions on which the conditional probabilities for the existence or the non-existence of an edge must be predicted. It should be noted that this is a very domain dependent task and would vary across networks and domains. In the experiments section we describe six combinations of networks for link prediction across networks. Due to limitations in space we describe the procedure for determining the conditions for just one of the cases.

Edges in the housing network in EQ2 can be of different types depending upon the type of relationship, corresponding to edge weight, that two players have with one another [1]. The strongest form of relationship is the Trustee relationship

between two nodes in the trust network and while the other types of relationship represent some form of linkage between two people, the relationship is not as strong. Given the nature of the housing network one would expect the Theory of Balance and the Theory of Homophily to play the most important part in explaining behavior in the housing network. While in the mentoring network the Theory of Cognition and the Theory of Self Interest would be the most prominent.

Thus one would expect the corresponding structures from Figure 1 to figure prominently in these networks. We get a more complex picture if we take all these factors together e.g., Theory of Homophily would predict that the links would be formed between nodes which are topologically closer together and have similar characteristics in the same network. The Theory of Self-Interest on the other hand implies that such structures are not going to be common. Taken together these observations imply that if a triangle is observed between these nodes in the housing network, it is unlikely to be present in the mentoring network between the same nodes. Another observation is the likelihood of formation of edges between nodes which are different with respect to expertise in the game. In the mentoring network this is going to be the case since mentoring relationship is established between actors if there is a difference in expertise between them. In the context of EQ2 this can be translated as the level of the player. However it is not possible to mentor any character in the game, they have to be in the same location at the same time.

Theory of Balance also implies that the more common neighbors that player have in a network, the more likely that they will form an edge. Additionally if the players have some other common identification e.g., guild membership then they are likely to form edges between themselves in case of the trust network and also mentor mentor other players in the guild if their level difference is sufficiently high.

V. Experiments

A. Dataset

Data from EQ2, a fantasy based MMORPG where thousands of players can simultaneously engage in many different types of activities like fighting non-player characters (NPCs), engaging in trade with other players, helping out other players, going on quests, exploring the landscape etc. Thus it is possible to construct multiple networks of players from this dataset. The game is played on multiple servers which can be thought of as 'parallel worlds.' Data from one of the servers 'guk' is used for experiments. We only consider the nodes in the largest connected component for our analysis. The following three networks are constructed from the game data:

- **Housing-Trust Network:** Player can then give access to other players to their virtual house. The social network is contructed on the basis of access ties.
- **Mentoring Network:** Players who are at a higher level can mentor players who are at a lower level and a social network is constructed based on this information.
- **Trade Network:** The social network constructed by creating an edge between two players if they engage in trade between one another.

Table II
Network Characteristics of EQ2 Networks

Net	N	E	d	NC	CC_1	CC_2
M	23,207	93,079	39	316	22,477	6
H	15,465	23,145	37	1,488	9,152	52
T	31,900	1,796,438	< 28	11	31,858	10

We use data from the game starting from January 1st, 2006 to September 11th, 2006. The global characteristics of these networks are given in Table II where N is the number of nodes, E is the number of edges, d is the diameter of the network and NC is the number of components. CC_1 and CC_2 are the largest and the second largest connected components respectively. M, H and T refer to the mentoring, housing and the trade network respectively.

We now describe the feature set which is used in the experiments. While the main focus of the paper is on topological features, we also include other types of features for comparison. Following is the list of features that we use in our experiments.

Proximity features: These are features that represent some form of proximity between a pair of nodes *e.g.,* two game characters may belong to the same guild/clan. The proximity features that are used here are defined in terms of indicator functions. If a_x is an attribute of node n_x then the indicator function can be given as:

$$s_{ij} = \begin{cases} 1, & \text{if } a_i = a_j \\ 0, & \text{if } a_i \neq a_j \end{cases} \quad (2)$$

The following indicator functions are used for the proximity featres: Real Gender Indicator, Real country indicator, Character class indicator, Character gender indicator, Character race indicator.

Aggregated features: These are combination of individual attributes of the node pair. The individual attribute can provide information which can help in the link prediction task *e.g.,* The higher the character level of a player, the more likely it is that it will interact with another character in some manner. Thus sum of character levels of a character pair can be a good aggregated feature. The following aggreated features are used: Sum of neighbors, Sum of actual age in 2006, Sum of joining age, Sum of character levels and Sum of character levels of the two game players.

Topological features: These are based on network topology. *e.g,* shortest distance between the pair of nodes. These are given below. The parametric versions of the common

neighbors, Adar-Adamic Index and Resource Allocation Index were given by [14]. Given nodes n_i and n_j, these features are defined as follows:

- Common Neighbors: If $\Gamma(x)$ represents the neighbors of x then:

$$s_{ij} = \Gamma(i) \cap \Gamma(j) \quad (3)$$

- Shortest distance: The shortest distance between the pair of nodes in the network.
- Clustering Coefficient: This is a measure of localized density and measures the participation of the nodes in triads.
- Adar Adamic Index: This metric improves on the common neighbors metric by giving more weight to the neighbors who are lower connecting nodes. If $f(k) = \Gamma(k)$:

$$s_{ij} = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{log f(k)} \quad (4)$$

- Resource Allocation Index: This metric is a modified form of the Adar-Adamic Index.

$$s_{ij} = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{f(k)} \quad (5)$$

- Parametric Weighted Common Neighbors: If $w(i,j) = w(j,i)$ is the weight of the links between nodes n_i and n_j then this metric is defined as follows. Note that if $\alpha = 0$ then the metric is equivalent is to the common neighbors metric and if $\alpha = 1$ then the metric is equal to taking the weights of the neighbors.

$$s_{ij} = \sum_{k \in \Gamma(i) \cap \Gamma(j)} w(i,k)^\alpha + w(k,j)^\alpha \quad (6)$$

- Parametric Adar Adamic: In this version of the Adar Adamic metric, 1 is added to $log(k)$ because the value of $s(k)$ may be less than 1 which may lead to negative values. The metric is given as follows:

$$s_{ij} = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{w(i,k)^\alpha + w(k,j)^\alpha}{log(1 + s(k))} \quad (7)$$

- Parametric Resource Allocation: The metric is given as follows:

$$s_{ij} = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{w(i,k)^\alpha + w(k,j)^\alpha}{s(k)} \quad (8)$$

While generalizations of the clustering coefficient exist, it has been noted [10] that the higher level analogues of the clustering coefficient are not really helpful in prediction.

B. Results

Given that there are thousands of nodes present in each network and thus millions of possible links between them, a prediction scheme which always predicts the non-existence of a link will get high precision and recall. To avoid this problem we randomly sample from the pairs of instances of nodes for positive and negative samples until we have the required number of examples, 60,000 in our case. We divide our data set into a training period spanning from January 2006 to June 2006 and test period spanning from July 2006 to September 2006. Following [4] the link prediction task in these experiments is defined as a machine learning problem where the binary classes are 'form-link' and 'do-not-form link.' A positive example is an edge in the test period which does not appear in the training period. A negative sample is an edge (with both of its nodes present in the training period) which is present neither in the training period nor in the test period. We used a total of 60,000 samples - with a maximum of 30,000 positive samples and the negative samples making up the rest. For validation and comparison with our technique we used six standard classification algorithms available in the popular machine learning library WEKA [8]. The algorithms that were used are J48, JRip, AdaBoost, Bayes Network, Nave Bayes and k-nearest neighbor, and 10-fold cross validation was used.

The problem of predicting across networks is non-trivial because participation of nodes in graph does not imply that these node are going to participate in other networks. This can be done by determining how much overlap there is between the various networks. For this purpose we computed the Adjacency Correlation, as defined by Clauset and Eagle [5] which determine the correlation between the adjacency matrices of two graphs.

$$\gamma_j = \frac{\sum_{i \in N_j} A^{(x)}_{i,j} A^{(y)}_{i,j}}{\sqrt{\left(\sum_{i \in N_j} A^{(x)}_{i,j}\right)\left(\sum_{i \in N_j} A^{(y)}_{i,j}\right)}} \quad (9)$$

Where $A(x)$ and $A(y)$ are the adjacency matrices of the graph at Time x and at time $y, N(j)$ is the union of row elements which are non-zero in at least one of the two matrices, γ_j is the correlation for the row for the two graphs. The adjacency correlation for the network is defined as the average of the adjacency correlation for all the rows in the adjacency matrix. The adjacency correlation between the three networks is given in Table III. The table indicates that there is very small overlap between the three networks which could partially explain why we are getting poor results for our predictors.

The adjacency correlation values from table III would seem to indicate that there is very little overlap between the various networks in terms of participation of nodes from one network to another network. The relationship between the networks is however more complex than this. To illustrate

Table III
ADJACENCY CORRELATION FOR THE NETWORKS

	Housing	Mentoring	Trade
Housing	1	0.10056	0.00669
Mentoring		1	0.00492
Trade			1

Table IV
JACCARD'S INDEX FOR THE NETWORKS

	Housing	Mentoring	Trade
Housing	1	0.34296	0.35019
Mentoring		1	0.55459
Trade			1

this we computed the Jaccard's Index for only the nodes in the networks without considering the edges. The Jaccard's index for the three networks is given in Table IV. From the table it is clear that there is a high degree of overlap between the networks especially in case of the trade and the mentoring network but from Table III we know that the adjacency correlation between these networks is low which would imply that although the same types of nodes are participating in these networks but in general they are not forming the same type of ties. This type of information should be included in future approaches to this problem in order to improve the results.

Results from Table V to X reveal that the proposed approach consistently performs better than the other approaches. The two instances where the performance of other approaches is comparable to the proposed approach is in the case where prediction has to be done on the Trade network. The reason for this is that the trade network is a very dense network which is evident from Table II which shows that there are more than 1.7 million edges in the trade network,

Table V
HOUSING TO MENTORING

Technique	Positive	Negative	Precision	Recall	F-Score
J48	4108	55892	0.741	0.225	0.345
JRip	4108	55892	0.751	0.167	0.273
AdaBoost	4108	55892	0.834	0.162	0.271
NaiveBayes	4108	55892	0.248	0.343	0.288
BayesNet	4108	55892	0.354	0.426	0.387
KNN	4108	55892	0.288	0.146	0.193
MINTP	4108	55892	0.458	0.398	0.426

Table VI
MENTORING TO HOUSING

Technique	Positive	Negative	Precision	Recall	F-Score
J48	2528	57472	0.696	0.284	0.404
JRip	2528	57472	0.667	0.313	0.426
NaiveBayes	2528	57472	0.173	0.257	0.207
BayesNet	2528	57472	0.278	0.464	0.348
AdaBoost	2528	57472	0.680	0.308	0.424
KNN	2528	57472	0.273	0.098	0.144
MINTP	2528	57472	0.581	0.398	0.472

Table VII
MENTORING TO TRADING

Technique	Positive	Negative	Precision	Recall	F-Score
J48	30001	29999	0.766	0.789	0.777
JRip	30001	29999	0.776	0.790	0.783
NaiveBayes	30001	29999	0.669	0.915	0.773
BayesNet	30001	29999	0.720	0.838	0.774
AdaBoost	30001	29999	0.790	0.746	0.767
KNN	30001	29999	0.736	0.748	0.742
MINTP	30001	29999	0.767	0.790	0.778

Table VIII
TRADING TO MENTORING

Technique	Positive	Negative	Precision	Recall	F-Score
J48	12740	47260	0.619	0.509	0.559
JRip	12740	47260	0.627	0.551	0.586
NaiveBayes	12740	47260	0.441	0.743	0.553
BayesNet	12740	47260	0.436	0.755	0.553
AdaBoost	12740	47260	0.593	0.563	0.578
KNN	12740	47260	0.545	0.483	0.512
MINTP	12740	47260	0.666	0.592	0.627

while in the other networks the number of links are less than a hundred thousand. Thus many nodes which are picked at random are likely to form an edge if they are sufficiently close to one another. In the other instances the proposed approach performs much better than other techniques.

VI. CONCLUSION

In this paper a new link prediction problem, inter-network link prediction, was proposed where the goal is to predict which links across multiple network are likely to form. Thus if information is available about node attributes and edge information from one network then one predict edges in another network where there is an overlap in membership between the two networks. While there are a large number of techniques for link prediction, these techniques seldom

Table IX
HOUSING TO TRADING

Technique	Positive	Negative	Precision	Recall	F-Score
J48	30001	29999	0.790	0.821	0.805
NaiveBayes	30001	29999	0.743	0.874	0.803
BayesNet	30001	29999	0.737	0.865	0.796
AdaBoost	30001	29999	0.809	0.788	0.798
KNN	30001	29999	0.769	0.785	0.777
MINTP	30001	29999	0.785	0.796	0.790

Table X
TRADING TO HOUSING

Technique	Positive	Negative	Precision	Recall	F-Score
J48	2869	57131	0.587	0.137	0.222
JRip	2869	57131	0.538	0.131	0.211
NaiveBayes	2869	57131	0.205	0.453	0.282
BayesNet	2869	57131	0.202	0.628	0.306
AdaBoost	2869	57131	0.538	0.005	0.010
KNN	2869	57131	0.363	0.163	0.225
MINTP	2869	57131	0.784	0.239	0.366

use knowledge from social science theories on network evolution. Data from a Massively Multiplayer Online Role Playing Game (MMORPG) EQ2 was used for experiments and validation. These networks are formed by different social processes and thus different feature set are helpful in making predictions in these networks. We then described a new technique which can be used for link prediction across networks which employs insights from social science theories to make predictions about links across networks. Specifically the MTML theory of Monge and Contractor [15] was employed to make predictions about the existence or non-existence of edges. Future work would involve extending the inter-network link prediction problem to include other characteristics of network and employ other datasets to test the generalizability of the proposed approaches.

ACKNOWLEDGMENT

The research reported herein was supported by the National Science Foundation via award number IIS-0729421, the Army Research Institute via award number W91WAW-08-C-0106, the Intelligence Advanced Research Projects Activity via AFRL Contract No. FA8650-10-C-7010, and the ARL Network Science CTA via BBN TECH/W911NF-09-2-0053. The data used for this research was provided by the SONY corporation. We gratefully acknowledge all our sponsors. The findings presented do not in any way represent, either directly or through implication, the policies of these organizations.

REFERENCES

[1] Muhammad Aurangzeb Ahmad, Marshall Scott Poole, Jaideep Srivastava, *Network Exchange in Trust Networks* IEEE Social Computing (SocialCom-10)

[2] Muhammad Aurangzeb Ahmad, David Huffakar, Annie Wang, Jeff Treem, Scott Poole, Jaideep Srivastava *GTPA: A Generative Model for Online Mentor-Apprentice Networks* Twenty-Fourth AAAI Conference on Artificial Intelligence Atlanta, Georgia July 11-15, 2010

[3] Adamic, L. A., Adar, E., *Friends and Neighbors on the Web*, Social Networks, Vol.25, No.3, pp.211-230, 2003.

[4] Al Hasan, M., Chaoji, V., Salem, S., Zaki, M. *Link prediction using supervised learning* Workshop on Link Analysis, Counter-terrorism and Security, SIAM, 2006.

[5] Clauset, A., Eagle, N. *Persistence and periodicity in a dynamic proximity network*, DIMACS Workshop on Comp. Methods for Dynamic Interaction Nets. (2007).

[6] Clauset, A., Moore, C., Newman, M. *Hierarchical structure and the prediction of missing links in networks*, Nature 453, 98-101 (2008).

[7] Contractor, N. S., Wasserman, S., Faust, K. *Testing multitheoretical multilevel hypotheses about organizational networks: An analytic framework & empirical example*. Academy of Manag. Review. (2006)

[8] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Ian, Witten *The WEKA Data Mining Software: An Update*; SIGKDD Explorations, Volume 11, Issue 1 (2009).

[9] Huang, Z., Li, X., Chen, H., *Link Prediction Approach to Collaborative Filtering*, 5th ACM/IEEECS joint conf. on Digital libraries, (2005)

[10] Huang, Z., *Link Prediction Based on Graph Topology: The Predictive Value of the Generalized Clustering Coefficient*, Workshop on Link Analysis: Dynamics and Static of Large Networks, the 12th ACM SIGKDD , Philadelphia, PA, 2006.

[11] Kunegis, J., Lommatzsch, A., *Learning spectral graph transformations for link prediction*, ICML (2009).

[12] Lee, D., Seung, H., *Algorithms for non-negative matrix factorization*. NIPS 13, 2001.

[13] David Liben-Nowell, Jon M. Kleinberg: The link prediction problem for social networks. CIKM 2003

[14] Lu, L., Zhou, T *Role of weak ties in link prediction of complex networks*. CIKM-CNIKM 2009: 55-58

[15] Monge, Peter., Contractor, N. *Theories of Communication Networks*. Cambridge: Oxford University Press (2003).

[16] Murata, Tsuyoshi., Moriyasu, Sakiko., *Link Prediction of Social Networks Based on Weighted Proximity Measures*. Web Intelligence 2007: 85-88

[17] Newman, M. E., *Clustering and Preferential Attachment in Growing Networks*, Physical Review Letters E, Vol.64 (025102), 2001.

[18] A. Potgieter, K. April, R. Cooke, I. O. Osunmakinde, *Temporality in Link Prediction: Understanding Social Complexity*, J. of Trans. on Eng. Management, (2006).

[19] Rattigan, Jensen. *The case for anomalous link discovery.* ACM SIGKDD, 2005.

[20] Sharan, U., Neville, J., *Exploiting Time-Varying Relationships in Statistical Relational Models*. SNA-KDD 2007.

[21] Tylenda, T., Angelova, R., Bedathur, S.,*Towards Time-aware Link Prediction in Evolving Social Networks*, KDD-SNA 2009

[22] Wasserman, Stanley., Faust, Katherine., *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press (1994).

[23] Watts, D., Strogatz, S., *Collective dynamics of 'small-world' networks*. Nature 393 40910. 1998.

[24] Williams, D., Ducheneaut, N., Xiong, L., Zhange, Y., Yee, N., Nickell, E. *From tree house to barracks: The social life of guilds in world of warcraft.* Games and Culture, 1, 338-363 (2006).

[25] Xiang, Evan W., *A Survey on Link Prediction Models for Social Network Data* PhD Qualifying Exam (2008).

[26] T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang. *Bipartite network projection and personal recommendation*. Phys. Rev. E, 76:046115, 2007.