Scaling up Link Prediction with Ensembles

Liang Duan[1], Charu Aggarwal[2], Shuai Ma[1], Renjun Hu[1], Jinpeng Huai[1]
[1]SKLSDE Lab, Beihang University, China [2]IBM T. J. Watson Research Center, USA
{duanliang, mashuai, hurenjun, huaijp}@buaa.edu.cn charu@us.ibm,com

Abstract

A network with n nodes contains $O(n^2)$ possible links. Even for networks of modest size, it is often difficult to evaluate all pairwise possibilities for links in a meaningful way. Hence, we propose an ensemble enabled approach to scaling up link prediction by decomposing traditional link prediction problems into subproblems of smaller size. These subproblems are each solved with the use of latent factor models, which can be effectively implemented over networks of modest size. Furthermore, the ensemble enabled approach has several advantages in terms of performance. Experiments on large networks demonstrate the effectiveness and scalability of our approach.

The $O(n^2)$ Problem in Link Prediction

Network Sizes	1 GHz	3 GHz	10 GHz
10^6 nodes	1000 sec.	333 sec.	100 sec.
10^7 nodes	27.8 hrs	9.3 hrs	2.78 hrs
10^8 nodes	> 100 days	> 35 days	> 10 days
10^9 nodes	> 10000 days	> 3500 days	> 1000 days

Table 1: Time required to allocate a single machine cycle to every node-pair possibility in networks.

Most current link prediction algorithms evaluate the link propensities only over a subset of possibilities rather than exhaustively search over the entire network [1, 2].

Latent Factor Model for Scalable Link Prediction

The link prediction ranking problem: Given a network $G = (N, A)$ with node set N and edge set A, determine the top-k node-pair recommendations such that these node pairs are not included in A.

We make a latent factor model, i.e., nonnegative matrix factorization (NMF), practical for link prediction by factorizing the weight matrix W of G into FF^T, where F is a set of latent factors and can be determined by using the following multiplicative update rule [3]:

$$F_{ij} \leftarrow F_{ij}(1-\beta+\beta\frac{(WF)_{ij}}{(FF^TF)_{ij}}), \beta \in (0,1]$$

The positive values of entries in FF^T can be viewed as the predictions of noisy 0-entries in W.

Efficient top-k prediction searching: We design an ε-approximate top-k method to speed up searching the k largest values in FF^T. The following nested loop is executed for each (say p-th) column of S:

for each i = 1 to f_p **do**
  for each $j = i + 1$ to f_p' **do**
    // S is obtained by sorting the columns of F in a descending order
    if $S_{ip} \cdot S_{jp} < \varepsilon / r$ **then** break inner loop;
    else increase the score of node-pair (R_{ip}, R_{jp}) by
        an amount of $S_{ip} \cdot S_{jp}$; // R is an inverted index matrix
  end for
end for

The complexity of NMF is $O(r\cdot(m + n\cdot r))$. For large sparse networks, the $O(nr^2)$ term might be the bottleneck.

Structural Bagging Methods

To scale up link prediction on very large networks, we develop three bagging methods:

(1) Select an ensemble component N_s form G by one of the node, edge and biased edge bagging methods;
(2) Construct a reduced adjacency matrix W_s from N_s;
(3) Apply NMF to W_s and use top-k method to predict.

If a node pair appears in multiple ensemble components, the maximum predicted value is considered.

Bound of node bagging ensembles: The expected times of each node pair included in μ / f^2 ensemble components is at least μ, where f is the sampling fraction.

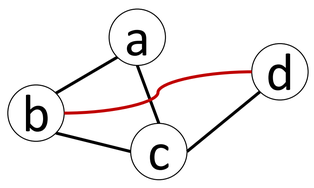
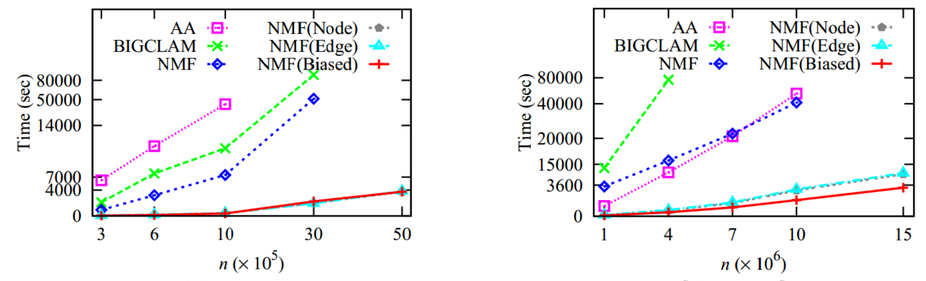


Figure 1: Triangle-closing model: The edge (b, d) is a triangle-closing edge. When the node c is selected, its neighbors a, b and d are also put into the same ensemble component.

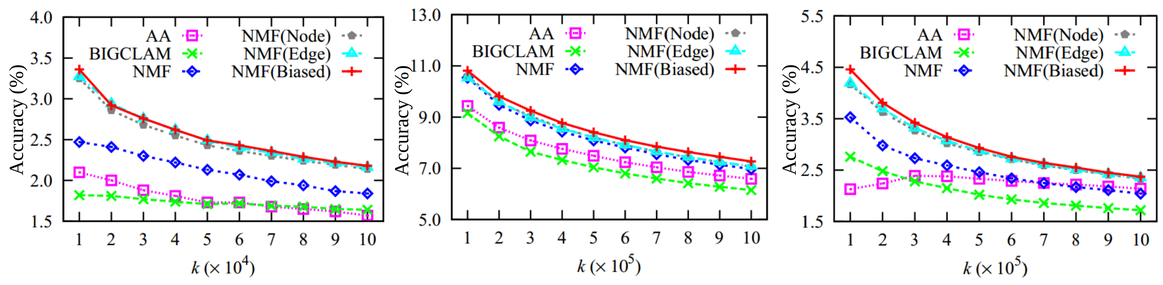
Using link prediction characteristics: Since most of all new links in social networks span within very short distances [4], typically closing triangles shown in Figure 1, we design particular bagging approaches such that a node is always sampled together with all its neighbors, which guarantees the possibility of forming triangles.

Experimental Results



(a) Twitter (b) Friendster

Figure 2: Efficiency comparison: w.r.t. the network sizes.



(a) YouTube (b) Flickr (c) Wikipedia

Figure 3: Accuracy comparison: w.r.t. the number k of predicted links.

Dataset	Accuracy	Dataset	Speedup
YouTube	18%	Twitter	20x
Flickr	4%		
Wikipedia	16%	Friendster	31x

Table 2: Summaries of the accuracy and efficiency improved by NMF(Biased) compared with NMF.

The ensemble-enabled approach not only increases the efficiency, but also improves the accuracy.

References

[1] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. *KDD*, 2011.

[2] C. Lee, M. Pham, N. Kim, M. K. Jeong, D. K. J. Lin and W. Art. A Novel Link Prediction Approach for Scale-free Networks. *WWW*, 2014.

[3] C. Ding, X. He and H. D. Simon. On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. *SDM*, 2005.

[4] J. Leskovec, L. Backstrom, R. Kumar and A. Tomkins. Microscopic Evolution of Social Networks. *KDD*, 2008.