Research Article

The Influence of Network Structural Preference on Link Prediction

Yongcheng Wang,1 Yu Wang,1 Xinye Lin,1 and Wei Wang2

1National Key Laboratory of Science and Technology on Blind Signal Processing, Chengdu 610041, China
2Cybersecurity Research Institute, Sichuan University, Chengdu 610065, China

Correspondence should be addressed to Wei Wang; wwzqbx@hotmail.com

Received 4 May 2020; Accepted 15 June 2020; Published 26 September 2020

Guest Editor: Chenquan Gan

Copyright © 2020 Yongcheng Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Link prediction in complex networks predicts the possibility of link generation between two nodes that have not been linked yet in the network, based on known network structure and attributes. It can be applied in various fields, such as friend recommendation in social networks and prediction of protein-protein interaction in biology. However, in the social network, link prediction may raise concerns about privacy and security, because, through link prediction algorithms, criminals can predict the friends of an account user and may even further discover private information such as the address and bank accounts. Therefore, it is urgent to develop a strategy to prevent being identified by link prediction algorithms and protect privacy, utilizing perturbation on network structure at a low cost, including changing and adding edges. This article mainly focuses on the influence of network structural preference perturbation through deletion on link prediction. According to a large number of experiments on the various real networks, edges between large-small degree nodes and medium-medium degree nodes have the most significant impact on the quality of link prediction.

1. Introduction

Complex networks play an important role in modeling and analyzing complex systems such as the social system, biological system, and information system [1]. Generally, individuals, such as human beings, biological elements, and computers, are represented by nodes, while links represent the relations or interaction between nodes [2]. The theory of complex networks offers new insight into the connection in the real world. Related research studies focus not only on single-layer networks, such as clustering [3], link prediction [4], and community discovery [5, 6] but also multilayer networks [7] as well, such as cascade [8, 9], communication [10–13], synchronization [14], and game [15–17].

Link prediction in a complex network includes prediction on links unknown or to be built in a given network. Based on the structure and attributes of a public network, link prediction is intended to predict the possibility of link generation between two nodes, which have not been connected yet [18]. For a given undirected network, as shown in Figure 1, the solid lines signify known edges in the network, while the dotted lines signify the unknown edges or those to be built in the future. What we need to do here is to predict the unknown edges denoted by the dotted lines through known nodes and edges accurately. As one of the research directions of data mining, link prediction can be applied to various fields. For instance, in social networks, link prediction can recommend new friends to users by predicting whether two strangers online are acquaintances offline. In biology, predicting protein-protein interaction will significantly reduce the cost of experiments. Besides, researchers have applied the idea and method of link prediction on node classification, such as determining the type of an article in an academic network [19].

Researchers have made considerable efforts to enhance the accuracy of link prediction. Many link prediction algorithms are based on the similarity between nodes, assuming that the more similar two nodes are, the more likely
that there is a link between them. The index describing the similarity between nodes can be roughly classified into the local index, global index, and quasi-local index. The local index is the most commonly used among the methods based on node similarity, due to its simplicity and adaptability for considerably large networks. There are other link prediction algorithms which could be based on maximum likelihood estimation, and they have better performance when dealing with networks with a distinct hierarchical structure, such as a grassland food chain [20]. Some algorithms may also employ probabilistic models for link prediction, during which information processed covers not only the network structure but also the attributes of nodes. These algorithms are characterized by higher prediction accuracy but, at the same time, greater calculation complexity [21].

Link prediction is usually proven useful in biological networks. In social networks, however, they may raise concerns about privacy and security, in that our data are valuable not only for enterprises and public entities but also for an increasing number of cybercriminals conducting network analysis for malicious purposes. Through the link prediction algorithm, cybercriminals can accurately predict the friends of a social account user and even the owner of that account according to his or her relationships. If they dig further, criminals may find the name, age, address, bank account, and other private information of a social account’s corresponding entity.

Considering what has been mentioned above, it is urgent to improve privacy protection. However, currently, there lacks intensive research on how to prevent identification of link prediction algorithms utilizing concealing, changing, or adding edges through network structural disturbance at a small cost. Based on the perturbation of the adjacency matrix, Lu et al. [22] studied the influence of structural consistency on link predictability. Waniek et al. [23] studied how to conceal sensitive relations in the network through reconnection strategy, proposing a heuristic method in achieving it. Wu et al. [24] proposed an active learning algorithm and applied perturbation on the most symbolic links in the network, to adjust the structure predictability of the graph.

Based on the previous research, this article focuses on the influence of network structural preference disturbance through deletion on link prediction. According to a large number of experiments on the various real networks, edges between large-small degree nodes and medium-medium degree nodes have the most significant impact on the performance of link prediction. In the real-world network, the connection choice between nodes is not uniform, but there is an obvious preference, which leads to a certain correlation between nodes in the network. Based on this connection correlation between nodes, people put forward the concept of homogeneity and heterogeneity to distinguish the connection preference between nodes. Therefore, the heterogeneity of complex network nodes is a measure of the uniform distribution of nodes. If the nodes tend to connect similar nodes, they will form homogeneous network; if the high nodes and low nodes have certain probability to connect, they will form heterogeneous network.

2. Model Demonstration

In this section, some basic terminologies used in this article will be first introduced, based on which official definitions will be made. Then, we will present the method of network structural preference perturbation. Finally, the pseudocode of this method will be given.

2.1. Definition. A complex network: a given biological or social network can be modeled as a graph, \( G = (V, E) \), in which \( V \) denotes the set of nodes in the network, \( E \) denotes the set of edges, and \( A \) denotes the adjacent matrix of the network. When there is a link \((i, j) \in E\) between node \(v_i\) and \(v_j\), the elements from matrix \(A\) satisfy \(e_{ij} = 1\); otherwise, they satisfy \(e_{ij} = 0\). We can use source and target to describe the relationship between nodes, thus dividing networks into directed and undirected networks. In directed networks, the link, by the name of arc, is built from a source node to a target node. In undirected networks, there is no distinction between source nodes and target nodes in a link, by the name of edge. This article will mainly focus on undirected networks, which can be further applied to directed networks conveniently.

Link prediction: in a given network \( G = (V, E) \), \( N \) denotes the number of nodes and \( M \) denotes the number of edges. Therefore, the total number of node pairs is \((N(N-1))/2\). If the universal set \(U\) denotes the set of all pairs of nodes in a network, link prediction works as follows: any pair of nodes, denoted by \((v_i, v_j) \in (U - E)\), which does not belong to the network, will be assigned a particular score through a certain kind of algorithm. Then, based on scores assigned to the pairs of nodes, the largest \( K \) pairs will be chosen as edges for prediction.

Network structural preference perturbation: in a given network \( G = (V, E) \), network structural preference perturbation aims at executing one or more operations among adding, changing, and deleting towards the edges of a network at a lower cost based on the set of edges \( E \). The aim of these operations is that when a new network \( G' = (V, E') \) is formed eventually, the edges lost or to be built in \( G \) are...
barely predictable through the operations demonstrated above.

2.2. Method of Perturbation. The method of network structural preference perturbation mainly consists of one or more operations among adding, changing, and deleting towards the edges of a network. This article will focus on deletion, trying to identify the particular quality of edges that are significant in influencing the effect of link prediction. For a given network $G = (V, E)$, the set of edges $E$ can be divided into a train set $E^T$ and test set $E^P$. The network structural preference perturbation mainly happens in train sets. After the perturbation, edge prediction can be made in the training network through an algorithm of link prediction. The performance of link prediction can thus be calculated by comparing the edges predicted and those in the test set. The given network in the following passages refers to a training network.

When a training network has been divided from the original one, we start to apply perturbation on it. For any edge denoted by $e_{ij}$, its deletion value can be calculated through the following formula:

$$w_{ij} = \frac{e_{ij} \cdot (k_i \cdot k_j)^{\alpha}}{\sum_{i=1}^{\left|V\right|} \sum_{d=1}^{\left|V\right|} e_{ab} \cdot (k_a \cdot k_b)^{\alpha}}.$$  

(1)

In this formula, $k_i$ and $k_j$ denote, respectively, the degree of node $v_i$ and $v_j$. $\alpha$ is used to control the preference of edge selection. When $\alpha$ is specified as a specific value, the weight of each edge selected will be calculated. For example, when $\alpha$ is slightly greater than 0, the connection between large and small nodes and medium and medium nodes has a larger weight, which means it is easier to be selected. $\alpha$ is an adjustable parameter. Formula (1) makes sure that the sum of deletion value of all the edges equals 1.

After acquiring the deletion value of every edge, we set up a parameter of proportion $f$, in order to randomly pick out edges by the number of $f \cdot |E^T|$ and delete them. It is noteworthy that the degree of some nodes in the edge may change after the edge has been deleted, and hence, the deletion value of every edge, $w_{ij}$, will be changed. In an ideal situation, the deletion value of the remaining edges should be recalculated every time an edge is deleted. However, the time complexity of the perturbation algorithm would be enormous in this way. To reduce time complexity, a parameter of the proportional interval is set by the name of $f_{\text{interval}}$, meaning that the calculation of deletion value is only redone after every $f_{\text{interval}} \cdot |E^T|$ edges are deleted.

The pseudocode of the proposed method is as follows. (see Algorithm 1).

3. Experiment

3.1. Experimental Setup. We have experimented on four real networks, whose statistics are shown in Table 1. During the experiment, four algorithms of link prediction are used, including RA [29], AA [30], CN’ [31], and PA [32]. For a random given node $v_i$, with $\Gamma(i)$ denoting the set of its adjacent nodes, the calculation formula of the four algorithms can be expressed as follows:

- Resource allocation (RA) is

$$s_{ij}^{RA} = \sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log k_z}.$$  

(2)

- Adamic-Adar (AA) index is

$$s_{ij}^{AA} = \sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\sum_z f}.$$  

(3)

- Common neighbor (CN) is

$$s_{ij}^{CN} = |\Gamma (i) \cap \Gamma (j)|.$$  

(4)

- Preferential attachment (PA) is

$$s_{ij}^{PA} = |k_i \cdot k_j|.$$  

(5)

We here choose precision as the index of the performance of link prediction. For a given group of edges that has not been observed, precision is defined as the ratio of successfully predicted edges to the top $L$ predicted edges. Suppose that $L_s$ is the number of successfully predicted edges among the top $L$ predicted edges; then, the calculation formula for precision can be expressed as follows:

$$\text{precision} = \frac{L_s}{L}.$$  

(6)

For a given network, the train set and test set will be divided by the ratio of 9: 1. In order to test the performance of prediction, perturbation will be applied on the train set under different $\alpha$ and $f$. Then, predictions will be made through the four algorithms of link prediction, and the result will be compared with the test set. Every experiment will be repeated 100 times.

3.2. Experiment Result. First, the experiment supposes that $f$ equals 0.01, 0.1, 0.2, 0.3, and 0.4, respectively, and $f_{\text{interval}} = 0.01$. The precision calculated through four algorithms on four data sets will be tested, under the condition that $\alpha$ ranges from $-20$ to $20$, with an interval of 1. As shown in Figure 2, the result indicates that when $\alpha$ is slightly larger than 0, the minimum value of precision is achieved, signifying the best effect of network perturbation. The result shown in the cases (a), (e), and (i) from Figure 3 suggests that edges between large-small degree nodes and medium-medium degree nodes have the largest influence on the performance of link prediction.

Then, $\alpha$ is set to be $-20$, $-10$, 0, 10, and 20, respectively, and $f_{\text{interval}} = 0.01$. The precision calculated through four algorithms on four data sets will be tested under the condition that $f$ ranges from 0.01 to 0.4, with an interval of 0.01. As shown in Figure 4, in most cases, the precision decreases as $f$ increases for all four methods of the algorithm. Because the larger the $f$ is, the larger the ratio of deleted edges will be, and the fewer edges are remained in the network, thus reducing the quality of prediction of the different algorithms of link prediction. However, there are cases in metabolic and...
(1) Input: adjacent matrix of the original network $A$, parameter $f$, parameter $f_{\text{interval}}$, and parameter $\alpha$

(2) for $i = 1; i \leq (f/f_{\text{interval}}) \times i + 1$ do

(3) $A = \text{DELETE-EDGE}(A, f_{\text{interval}}, \alpha)$

(4) end for

(5) function $\text{DELETE-EDGE}(A, f_{\text{interval}}, \alpha)$

(6) $|E| = \text{sum}(\text{sum}(A))$

(7) Initialize the value matrix: $W \leftarrow 0$

(8) for $i = 1; i \leq \text{size}(A) ; i + 1$ do

(9) for $j = 1; j \leq \text{size}(A) ; j + 1$ do

(10) Calculate the value of each element $w_{ij}$ in $W$ using formula (1);

(11) end for

(12) end for

(13) for $k = 1; k \leq f_{\text{interval}} \times |E| ; k + 1$ do

(14) Randomly choose a position $(i, j)$ based on the value matrix $W$;

(15) $A_{ij} \leftarrow 0$

(16) end for

(17) return $A$

(18) end function

(19) Output: the adjacent matrix after the perturbation $A^* = A$ has been applied

**Algorithm 1:** Perturbation algorithm.

**Table 1:** Statistical features of four real networks, including the number nodes $|V|$, the number of edges $|E|$, average degree $\langle k \rangle$, average shortest path $\langle d \rangle$, clustering coefficient $C$ [28], assortativity coefficient $r$ [33], and heterogeneity of node degree $H (H = (\langle k^2 \rangle/\langle k \rangle^2))$.

| Networks     | $|V|$ | $|E|$ | $\langle k \rangle$ | $\langle d \rangle$ | $C$  | $r$  | $H$   |
|--------------|------|------|---------------------|---------------------|------|------|-------|
| Jazz [25]    | 198  | 2,742| 27.697              | 2.235               | 0.618| 0.020| 1.396 |
| Macaques [26]| 62   | 1,187| 38.290              | 1.380               | 0.667| -0.073| 1.039 |
| Metabolic [27]| 453 | 2,025| 8.940               | 2.664               | 0.647| -0.226| 4.485 |
| Neural [28]  | 297  | 2,148| 14.465              | 2.455               | 0.292| -0.163| 1.801 |

**Figure 2:** Continued.
Figure 2: Precision under different $\alpha$. From the top to bottom, results are shown through four methods of RA, AA, CN, and PA. From left to right, results are shown from four networks of jazz, macaque, metabolic, and neural.

Figure 3: Analysis of the reason why perturbation effect reaches its peak when $\alpha$ is slightly larger than 0. (a) Connection of large-small degree nodes; (b) connection of large-medium degree nodes (c) connection of large-large degree nodes; (d) connection of medium-small degree nodes; (e) connection of medium-medium degree nodes; (f) connection of medium-large degree nodes; (g) connection of small-small degree nodes; (h) connection of small-medium degree nodes; (i) connection of small-large degree nodes.
Figure 4: Precision under the condition of various $f$. From the top to bottom, results are shown through four methods of RA, AA, CN, and PA. From left to right, results are shown from four networks of jazz, macaque, metabolic, and neural.
neural networks where precision increases as $f$ increases, which could result from the fact that metabolic and neural networks have higher heterogeneity of node degree.

In order to better demonstrate the influence network structural preference perturbation has on link prediction, we have also tested the precision calculated through four algorithms on four datasets, under the condition that $\alpha$ ranges from $-20$ to $20$, with an interval of $1$, $f$ ranges from $0.01$ to $0.4$, with an interval of $0.01$, and $f_{\text{interval}} = 0.01$. The experimental results are shown in Figure 5, where curves in the horizontal planes are isolines.

4. Conclusion

In this article, the influence of network structural preference perturbation by a deletion on link prediction is analyzed. By using an interactive criterion to determine node degree, we first assign a perturbation value through the calculation to
every edge in a given network. Then, we apply perturbation through deletion on edges selected according to perturbation value. This procedure will be repeated until a certain proportion of edges have experienced perturbation. After that, we make link prediction on networks before and after perturbation, using four methods including RA, AA, CN, and PA, compared to the different influence types of connection and the ratio of deletion has on the performance of link prediction.

Massive experiments on various real networks indicate that the edges between large-small degree nodes and those between medium-medium degree nodes have the most significant influence on the performance of link prediction. By deleting the specific link in the network, we can resist the impact of link prediction on privacy protection. The above strategies can not only protect privacy in the field of social networks but also be worth promoting and applying in other fields. For example, in the design of computer communication topology, to minimize the connection between large and small nodes, medium and medium nodes can resist topology estimation, so as to better protect our own network; in the field of counter-terrorism, we should pay more attention to the connection between the leader node and leaf node, which often means the vulnerability of the terrorist team in communication connection.

Data Availability
The data can be obtained upon request to the corresponding author.

Conflicts of Interest
The authors declare that they have no conflicts of interest.

Acknowledgments
This work was partially supported by the National Natural Science Foundation of China (Grant no. 61903266), China Postdoctoral Science Foundation (Grant no. 2018M631073), China Postdoctoral Science Special Foundation (Grant no. 2019T120829), Fundamental Research Funds for the Central Universities, and Sichuan Science and Technology Program (No. 20YYJ(C)4001).

References
[1] R. Albert and A.-L. Barabási, “Statistical mechanics of complex networks,” Reviews of Modern Physics, vol. 74, no. 1, pp. 47–97, 2002.
[2] M. Newman, Networks: An Introduction, Oxford University Press, New York, NY, USA, 2010.
[3] X. Han, L. Wang, C. Cui, J. Ma, and S. Zhang, “Linking multiple online identities in criminal investigations: a spectral co-clustering framework,” IEEE Transactions on Information Forensics and Security, vol. 12, no. 9, pp. 2242–2255, 2017.
[4] K. Chi, G. Yin, Y. Dong, and H. Dong, “Link prediction in dynamic networks based on the attraction force between nodes,” Knowledge-Based Systems, vol. 181, 2019.
[5] G. Yin, K. Chi, Y. Dong, and H. Dong, “An approach of community evolution based on gravitational relationship refactoring in dynamic networks,” Physics Letters A, vol. 381, no. 16, pp. 1349–1355, 2017.
[6] Z. Zhao, C. Li, X. Zhang, F. Chiclana, and E. H. Viedma, “An incremental method to detect communities in dynamic evolving social networks,” Knowledge-Based Systems, vol. 163, pp. 404–415, 2019.
[7] R. Tang, S. Jiang, X. Chen, H. Wang, W. Wang, and W. Wang, “Interlayer link prediction in multiplex social networks: an iterative degree penalty algorithm,” Knowledge-Based Systems, vol. 194, p. 105598, 2020.
[8] J. Gao, S. V. Buldyrev, H. E. Stanley, and S. Havlin, “Networks formed from interdependent networks,” Nature Physics, vol. 8, no. 1, pp. 40–48, 2011.
[9] J. Gao, S. V. Buldyrev, S. Havlin, and H. E. Stanley, “Robustness of a network of networks,” Physical Review Letters, vol. 107, no. 19, 2011.
[10] C. Granell, S. Gómez, and A. Arenas, “Dynamical interplay between awareness and epidemic spreading in multiplex networks,” Physical Review Letters, vol. 111, no. 12, p. 128701, 2013.
[11] W. Wang, M. Tang, H. Yang, Y. Younghae Do, Y.-C. Lai, and G. Lee, “Asymmetrically interacting spreading dynamics on complex layered networks,” Scientific Reports, vol. 4, no. 1, p. 5097, 2014.
[12] W. Wang, Q.-H. Liu, S.-M. Cai, M. Tang, L. A. Braunstein, and H. E. Stanley, “Suppressing disease spreading by using information diffusion on multiplex networks,” Scientific Reports, vol. 6, no. 1, Article ID 29259, 2016.
[13] W. Wang, Q.-H. Liu, J. Liang, Y. Hu, and T. Zhou, “Co-evolution spreading in complex networks,” Ithaca, NY, USA, Physics Reports.
[14] X. Zhang, S. Boccaletti, S. Guan, and Z. Liu, “Explosive synchronization in adaptive and multilayer networks,” Physical Review Letters, vol. 114, no. 3, Article ID 038701, 2015.
[15] Z. Wang, A. Szolnoki, and M. Perc, “Optimal interdependence between networks for the evolution of cooperation,” Scientific Reports, vol. 3, no. 1, p. 2470, 2013.
[16] Z. Wang, A. Szolnoki, and M. Perc, “Interdependent network reciprocity in evolutionary games,” Scientific Reports, vol. 3, p. 1183, 2013.
[17] Z. Wang, A. Szolnoki, and M. Perc, “Self-organization towards optimally interdependent networks by means of co-evolution,” New Journal of Physics, vol. 16, no. 3, Article ID 033041, 2014.
[18] vol. 39, no. 5, pp. 651–661, 2010.
[19] B. Gallagher, H. Tong, T. Eliassi-Rad, and C. Faloutsos, “Using ghost edges for classification in sparsely labeled networks,” in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 256–264, Las Vegas, NV, USA, August 2008.
[20] A. Clauset, C. Moore, and M. E. J. Newman, “Hierarchical structure and the prediction of missing links in networks,” Nature, vol. 453, no. 7191, pp. 98–101, 2008.
[21] B. Taskar, M.-F. Wong, P. Abbeel, and D. Koller, “Link prediction in relational data,” Advances in Neural Information Processing Systems, pp. 659–666, 2004.
[22] L. Li, T. Zhou, Y.-C. Zhang, and H. E. Stanley, “Toward link predictability of complex networks,” Proceedings of the National Academy of Sciences, vol. 112, no. 8, pp. 2325–2330, 2015.
[23] M. Waniek, K. Zhou, Y. Vorobeychik, E. Moro, T. P. Michalak, and T. Rahwan, “How to hide ones relationships from link
prediction algorithms,” *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[24] T. Wu, G. Ming, X. Xian, W. Wang, S. Qiao, and G. Xu, “Structural predictability optimization against inference attacks in data publishing,” *IEEE Access*, vol. 7, pp. 92119–92136, 2019.

[25] P. M. Gleiser and L. Danon, “Community structure in jazz,” *Advances in Complex Systems*, vol. 06, no. 4, pp. 565–573, 2003.

[26] Y. Takahata, *Diachronic Changes in the Dominance Relations of Adult Female Japanese Monkeys of the Arashiyama B Group, the Monkeys of Arashiyama*, pp. 123–139, State University of New York Press, New York, NY, USA, 1991.

[27] J. Duch and A. Arenas, “Community detection in complex networks using extremal optimization,” *Physical Review E*, vol. 72, no. 2, Article ID 027104, 2005.

[28] D. J. Watts and S. H. Strogatz, “Collective dynamics of “small-world” networks,” *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[29] Q. Ou, Y.-D. Jin, T. Zhou, B.-H. Wang, and B.-Q. Yin, “Power-law strength-degree correlation from resource-allocation dynamics on weighted networks,” *Physical Review E*, vol. 75, no. 2, Article ID 021102, 2007.

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

[31] M. E. Newman, “Clustering and preferential attachment in growing networks,” *Physical Review E*, vol. 64, no. 2, Article ID 026102, 2001.

[32] L. Lü and L. Lü, “Link prediction in complex networks: a survey,” *Physica A: Statistical Mechanics and Its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

[33] M. E. Newman, “Assortative mixing in networks,” *Physical Review Letters*, vol. 89, no. 20, Article ID 208701, 2002.