A mathematical model for dynamics of occurrence probability of missing links in predicted missing link list

WenJun Zhang
School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China; International Academy of Ecology and Environmental Sciences, Hong Kong
E-mail: zhwj@mail.sysu.edu.cn, wjzhang@iaees.org

Received 25 December 2015; Accepted 15 February 2016; Published 1 December 2016

Abstract
In most of the link prediction methods, all predicted missing links are ranked according to their scores. In the practical application of prediction results, starting from the first link that has the highest score in the ranking list, we verify each link one by one through experiments or other ways. Nevertheless, how to find an occurrence pattern of true missing links in the ranking list has seldomly reported. In present study, I proposed a mathematical model for relationship between cumulative number of predicted true missing links (y) and cumulative number of predicted missing links (x): y = K(1-e^{-rx/K}), where K is the expected total number of true missing links, and r is the intrinsic (maximum) occurrence probability of true missing links. It can be used to predict the changes of occurrence probability of true missing links, assess the effectiveness of a prediction method, and help find the mechanism of link missing in the network. The model was validated by six prediction methods using the data of tumor pathways.

Keywords mathematical model; missing links; prediction; occurrence probability; tumor pathways.

1 Introduction
Link prediction is conducted to estimate the likelihood of the existence of a link between two nodes based on observed links and (or) the attributes of nodes (Zhang, 2015d; Zhou, 2015), which is expected to reduce the experimental costs for link finding. So far, numerous research on link prediction have been conducted (Clauset et al., 2008; Guimera and Sales-Pardo, 2009; Lü and Zhou, 2011; Lü et al., 2012; Barzel and Barabási, 2013; Bastiaens et al., 2015; Lü et al., 2015; Zhang, 2007, 2011, 2012a-c, 2013, 2015a-d, 2016a-d; Zhang and Li, 2015; Zhao et al., 2015; Zhou, 2015). Known so many prediction methods, each of them has its own mechanism for link generation, and their predictions for the same dataset are diverse. For example, the random perturbation method of Zhang (2016a) was developed based on degree growth of scale-free. Power-law generation method assumed that the degree of most networks is power-law distributed and power-law
distribution was used to fit degree distribution and predict missing links (Zhang, 2016c). CN was based on common neighbors of two links (Lorrain and White, 1971). In most of the link prediction methods, all predicted missing links are ranked based on their scores. In the application of prediction results, starting from the first link, we usually verify each link through experiments or other ways. How to find an occurrence pattern of true missing links in the ranking list for reducing verification cost has seldomly reported. In present study, I tried to propose a mathematical model for relationship between cumulative number of predicted true missing links and cumulative number of predicted missing links in order to further enhance prediction efficiency and identify the mechanism of link missing.

2 Methods
2.1 Mathematical model

In the most of prediction methods for missing links, for a link being predicted, calculate the score (likelihood, probability, etc. A higher score means the greater probability of being a true missing link) of the link using the prediction method, and rank all predicted missing links according to their scores, from greater to smaller ones. The top links in the ranking list are more likely true missing links. The ranking list is always stored in a file with three columns, A, B and C. A stores IDs of “from” nodes, B stores IDs of “to” nodes, and C stores scores of predicted missing links.

First, define the occurrence probability of true missing links as

\[ p = \frac{dy}{dx} \]  \hspace{1cm} (1)

where \( y \) is the cumulative number of predicted true missing links, and \( x \) is the cumulative number of predicted missing links, starting from the 1st predicted link with the highest score in the ranking list. A preceding link in the ranking list is more likely a true missing link than its succedent links. As a consequence, the occurrence probability of true missing links, \( p \), will decline (from the intrinsic occurrence probability, i.e., maximum occurrence probability, \( r \)) to zero as the increase of cumulative number of predicted true missing links \( y \) until the expected total number of true missing links, \( K \), is achieved. Therefore, as the first-order approximation of equation (1), let

\[ p = r(K - y)/K \] \hspace{1cm} (2)

and we have

\[ dy/dx = r(K - y)/K \] \hspace{1cm} (3)

where \( K > 0, 0 < r \leq 1 \). Solving equation (3), the mathematical model for relationship between cumulative number of predicted true missing links \( y \) and cumulative number of predicted missing links \( x \) is achieved as

\[ y = K(1 - e^{-rx/K}) \] \hspace{1cm} (4)

According to the model (4), the cumulative number of predicted true missing links \( y \) increases as the increase of predicted missing links \( x \) in the ranking list, and tends to an asymptote, i.e., expected total number of true missing links, \( K \) (Fig. 1).

Model (4) can be used to predict the changes of occurrence probability of true missing links in the ranking...
list. Based on the model (4), the expected total number of true missing links, $K$, and the intrinsic occurrence probability of true missing links, $r$, can be obtained by using data fitting.

The following are Matlab codes, linkPredModel.m, to obtain the relationship between $y$ and $x$ and the numerical method to obtain $K$ and $r$ by fitting relationship between $y$ and $x$.

```matlab
scores=input('input the excel file name of scores: ','s');
mislinks=input('input the excel file name of missing links: ','s');
scores=xlsread(scores); mislinks=xlsread(mislinks);
n=size(scores,1); m=size(mislinks,1);
scores=sortrows(scores,-3);
x=1:n;
y=zeros(1,n);
ma=0;
for i=1:n
    for j=1:m
        if ((scores(i,1)==mislinks(j,1)) & (scores(i,2)==mislinks(j,2))) y(i)=ma+1; ma=y(i);break; end
        if ((scores(i,1)==mislinks(j,2)) & (scores(i,2)==mislinks(j,1))) y(i)=ma+1; ma=y(i); break; end
    end
    y(i)=ma;
end
plot(x,y,'-');
xlabel('Cumulative number of predicted missing links (x)');
ylabel('Cumulative number of predicted true missing links (y)');
disp('Cumu. number of predi. missing links (x)      Cumu. number of predi. true missing links (y)')
[x'   y']
k=input('Input the estimated value of parameter K (e.g., 15): ');
r=input('Input the estimated value of parameter r (e.g., 0.1): ');
sig=input('Input the significance level (e.g., 0.01): ');
beta=[k r];
[beta,RJ,SIGMA,MSE]=nlmefit(x,y,@predictfunction,beta);
K=beta(1)
r=beta(2)
deltabeta=nlparci(beta,RJ);
fitted=predictfunction(beta,x);
chi_square=sum((y-fitted).^2./fitted)
p=chi2cdf(chi_square,n-2)
if (p<sig) disp('The data fit model well at the given significance level.');
else disp('The data is not able to fit model at the given significance level.');
end

The following is the function predictFunction.m

function f=predictfunction(beta,x)
f=beta(1)*(1-exp(-beta(2)/beta(1)*x));
```

The following is the function predictFunction.m

```matlab
function f=predictfunction(beta,x)
f=beta(1)*(1-exp(-beta(2)/beta(1)*x));
```
The software and data can be found in supplementary material of the present article.

![Graph](https://via.placeholder.com/150)

**Fig. 1** Theoretical relationship between predicted true missing links and predicted missing links in the ranking list.

### 2.2 Implication and significance of model (4)

Suppose there are totally \( n \) predicted missing links in the ranking list. With the same total number \( (S) \) of predicted true missing links in the ranking list, the method with the larger value of \( r \) is a better method (correspondingly, with the same value of \( r \), the method with the larger \( S \) is a better method). Suppose the total number of true missing links, which is unknown, is \( L \). Obviously, \( K \) is the approximation of \( L \), and \( S \leq L \). As \( r \rightarrow 1 \), and \( S \rightarrow L \), the prediction method tends to be a better method. Theoretically, \( r \) and \( K \) can be jointly used to assess performance of prediction methods. However, the estimated \( K \) may not well approximate \( L \) due to data quality for parameter fitting, and \( S \) is unknown at the early stage of link verification. Therefore, the parameter, \( r \), is more reliable for use. In summary, main implication and application of model (4) include

1. Intrinsic occurrence probability of true missing links \( (r) \) denotes the prediction capability or effectiveness of a prediction method and can thus be used to assess the performance of a prediction method.
2. It is naturally concluded that the mechanism of link missing for the better method in the assessment is more likely the true mechanism of link missing.
3. Model (4) (exactly, equation (2)) can be used to predict or estimate the occurrence probability of true missing links.

### 2.3 Prediction methods

In present study, six methods are used to predict missing links: (1) Random perturbation (Zhang, 2016a); (2) Power-law generation (Zhang, 2016c); (3) CN (common neighbors) (Lorrain and White, 1971); (4) Adamic-Adar (Adamic LA, Adar E. 2003); (5) RA (resource allocation; Zhou et al., 2009), and (6) Katz (Katz, 1953).

I use the data of original tumor pathways FAS, JAK-STAT, JNK, MARK and p53 (ABCAM, 2012; Huang and Zhang, 2012; Li and Zhang, 2013; Pathway Central, 2012; Zhang, 2016a; *pathwayname.xls* in
supplementary material). Links are removed from tumor pathways and the remaining (see missinglinks\_pathname.xls and \_Training.xls in supplementary material) is used to generate the training adjacency matrix.

Missing links are removed from original pathways following the inverse evolution process of the networks with power-law degree distribution. Therefore, the links missed following a known mechanism and the power-law based methods (Zhang, 2016c), in particular, random perturbation (Zhang, 2016a) are the methods featured by this mechanism. Most networks have the degree distribution of power-law (Barabasi and Albert, 1999; Zhang, 2012a, 2016a, 2016c). For network evolution based link prediction, it is a reasonable and exact treatment.

The prediction results of missing links in tumor pathways, by six methods, can be found in separate directories of supplementary material (PrediScores\_\_methodname.xls).

3 Results

3.1 Validation of the mathematical model

Compared with Fig. 1 and Fig. 2, a general coincidence between theoretical curve and observed curves can be founded. As indicated in Table 1, parameter fitting and statistic test further validate the mathematical model (4) that describes the relationship between cumulative number of predicted true missing links ($y$) and cumulative number of predicted missing links ($x$). In addition, the results show that the parameter $r$ is more reliable than $K$.

<table>
<thead>
<tr>
<th>Pathways</th>
<th>Model para./Chi square</th>
<th>Random perturbation</th>
<th>Power-law generation</th>
<th>CN</th>
<th>AA</th>
<th>RA</th>
<th>Katz</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAS ($L=12$ true missing links)</td>
<td>$K$</td>
<td>9.2480</td>
<td>28.2614</td>
<td>$-2.6\times10^9$</td>
<td>$1.2\times10^9$</td>
<td>$1.2\times10^9$</td>
<td>4.0375</td>
</tr>
<tr>
<td></td>
<td>$r$</td>
<td>0.0458</td>
<td>0.0177</td>
<td>0.0141</td>
<td>0.0142</td>
<td>0.0142</td>
<td>0.0190</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>330.5661**</td>
<td>92.2104**</td>
<td>57.1607**</td>
<td>169.0797</td>
<td>169.0797</td>
<td>61.7429**</td>
</tr>
<tr>
<td>JAK-STAT ($L=11$ true missing links)</td>
<td>$K$</td>
<td>7.2210</td>
<td>13.9050</td>
<td>3.8933</td>
<td>4.1197</td>
<td>4.1197</td>
<td>5.2$\times10^6$</td>
</tr>
<tr>
<td></td>
<td>$r$</td>
<td>0.0158</td>
<td>0.0250</td>
<td>0.2368</td>
<td>0.2042</td>
<td>0.2042</td>
<td>0.0070</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>36.1173**</td>
<td>54.3138**</td>
<td>5.3417**</td>
<td>20.6666**</td>
<td>20.6666**</td>
<td>139.9072**</td>
</tr>
<tr>
<td>JNK ($L=16$ true missing links)</td>
<td>$K$</td>
<td>16.0194</td>
<td>21.8450</td>
<td>2.2015</td>
<td>4.4009</td>
<td>5.5020</td>
<td>6.8$\times10^5$</td>
</tr>
<tr>
<td></td>
<td>$r$</td>
<td>0.1672</td>
<td>0.0347</td>
<td>0.0359</td>
<td>0.0147</td>
<td>0.0138</td>
<td>0.0172</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>43.4027**</td>
<td>127.4522**</td>
<td>12.1910**</td>
<td>21.7042**</td>
<td>21.3651**</td>
<td>695.3245**</td>
</tr>
<tr>
<td>MARK ($L=15$ true missing links)</td>
<td>$K$</td>
<td>3.2$\times10^3$</td>
<td>4.9$\times10^9$</td>
<td>2.2247</td>
<td>2.1745</td>
<td>2.1745</td>
<td>15.7222</td>
</tr>
<tr>
<td></td>
<td>$r$</td>
<td>0.0104</td>
<td>0.0119</td>
<td>0.0460</td>
<td>0.0296</td>
<td>0.0296</td>
<td>0.0122</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>117.8132**</td>
<td>71.5835**</td>
<td>36.0525**</td>
<td>69.4835**</td>
<td>69.4835**</td>
<td>496.0534**</td>
</tr>
<tr>
<td>p53 ($L=13$ true missing links)</td>
<td>$K$</td>
<td>11.2094</td>
<td>15.0853</td>
<td>1.7$\times10^{10}$</td>
<td>3.0554</td>
<td>3.0530</td>
<td>6.8874</td>
</tr>
<tr>
<td></td>
<td>$r$</td>
<td>0.0788</td>
<td>0.0163</td>
<td>0.0039</td>
<td>0.2080</td>
<td>0.2250</td>
<td>0.0327</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>59.8453**</td>
<td>653.3348**</td>
<td>56.2755**</td>
<td>6.8479**</td>
<td>7.4530**</td>
<td>26.0670**</td>
</tr>
</tbody>
</table>

$**: p<0.01; *: p<0.05.$
Fig. 2 Fitted and observed relationships based on model (4) (●●●: observed; ▬: fitted).

Table 2 Summary of prediction results of six methods for missing links in five tumor pathways.

<table>
<thead>
<tr>
<th>Pathways</th>
<th>Type of links</th>
<th>Random perturbation</th>
<th>Power-law generation</th>
<th>CN</th>
<th>AA</th>
<th>RA</th>
<th>Katz</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAS</td>
<td>Predicted missing links (n)</td>
<td>640</td>
<td>866</td>
<td>105</td>
<td>105</td>
<td>105</td>
<td>386</td>
</tr>
<tr>
<td></td>
<td>Predicted true missing links (S)</td>
<td>9</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>JAK-STAT</td>
<td>Predicted missing links (n)</td>
<td>545</td>
<td>613</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>736</td>
</tr>
<tr>
<td></td>
<td>Predicted true missing links (S)</td>
<td>5</td>
<td>9</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>JNK</td>
<td>Predicted missing links (n)</td>
<td>612</td>
<td>709</td>
<td>201</td>
<td>201</td>
<td>201</td>
<td>759</td>
</tr>
<tr>
<td></td>
<td>Predicted true missing links (S)</td>
<td>16</td>
<td>16</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>MARK</td>
<td>Predicted missing links (n)</td>
<td>819</td>
<td>1006</td>
<td>133</td>
<td>133</td>
<td>133</td>
<td>1166</td>
</tr>
<tr>
<td></td>
<td>Predicted true missing links (S)</td>
<td>8</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>p53</td>
<td>Predicted missing links (n)</td>
<td>539</td>
<td>1067</td>
<td>142</td>
<td>142</td>
<td>142</td>
<td>406</td>
</tr>
<tr>
<td></td>
<td>Predicted true missing link (S)</td>
<td>12</td>
<td>13</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>
Fig. 3 Comparisons of relationships between six methods under five tumor pathways (●●●: Power-law generation; ▲▲▲: Random perturbation; ■■■: Katz; ····: CN; ⋯⋯: RA; ──: AA). For FAS and JNK prediction, Random perturbation is significantly better than other methods, with the larger \( r \) values (steeper and earlier sprouted curves) and overall the larger total number of predicted true missing links than other methods, in exception of Power-law generation.

3.2 Judgement of prediction methods and discovery of link missing mechanism

A comprehensive survey on Table 1, Table 2 and Fig. 2, in terms of \( r \) and total number of true missing links \( S \), proves that Random perturbation performs significantly better than other methods, seconded by CN. The curves of Random perturbation (Fig. 2) are more similar to the theoretical curve in Fig. 1. The mechanism of link missing in Random perturbation is forseeablely the true mechanism.
Discussion
I guess that the pattern described by model (4) may have resulted from power-law distribution of true missing links in the ranking list, i.e., the occurrence probability of true missing links in the ranking list follows the power-law distribution.

Most of prediction methods were based on static topological structure only. Network evolution based (Zhang, 2012a, 2012c, 2015a, 2016b, 2016c), node similarity based (Zhang, 2015d), and sampling based (correlation based; Zhang, 2007, 2011, 2012b, 2013, 2015b; Zhang and Li, 2015) methods are used also. Model (4) in present study is suitable to predictions of all these types of methods.

To simply obtain $K$, we may carefully choose three points, $(x_1, y_1)$, $(x_2, y_2)$, $(x_3, y_3)$, where $x_3-x_2=x_2-x_1$. Derived from equation (3), we have the estimate of $K$ as the following

$$K= \frac{(y_2-y_1)(y_2-y_3)(2y_2-y_1-y_3)}{2(y_2-y_1+y_3)}$$

However, a sound method to obtain $K$ and $r$ from data fitting is urgently needed.

Acknowledgment
We are thankful to the support of Discovery and Crucial Node Analysis of Important Biological and Social Networks (2015.6-2020.6), from Yangling Institute of Modern Agricultural Standardization, China.

References