Article

A Fusion Link Prediction Method Based on Limit Theorem

Yiteng Wu *, Hongtao Yu*, Ruiyang Huang, Yingle Li and Senjie Lin

National Digital Switching System Engineering and Technological R&D Center, Zhengzhou 450002, China
* Correspondence: wuyiteng1992@163.com (Y.W.); 15937101921@139.com (H.Y.); Tel.: +86-150-9348-9229

Abstract: The theoretical limit of link prediction is a fundamental problem in this field. Taking the network structure as object to research this problem is the mainstream method. This paper proposes a new viewpoint that link prediction methods can be divided into single or combination methods, based on the way they derive the similarity matrix, and investigates whether there a theoretical limit exists for combination methods. We propose and prove necessary and sufficient conditions for the combination method to reach the theoretical limit. The limit theorem reveals the essence of combination method that is to estimate probability density functions of existing links and nonexistent links. Based on limit theorem, a new combination method, theoretical limit fusion (TLF) method, is proposed. Simulations and experiments on real networks demonstrated that TLF method can achieve higher prediction accuracy.

Keywords: link prediction; combination method; theoretical limit; TLF method

1. Introduction

Limit theory is a basic theoretical issue and has attracted wide interest across many fields. On the 100th anniversary of its foundation, *Science* raised 125 unresolved scientific questions, and many of these issues related to limit theory [1]. Link prediction predicts missing links in current networks and new or dissolution links in future networks [2]. With continuous improvement of link prediction methods and, the theoretical limit of link prediction has attracted considerable research interest [3].

Considering structure or attribute features, link prediction methods based on classification have been proposed by computer science community [4]. Subsequently, more insightful methods of network structure, such as similarity based methods, have become a focus, these methods pay more attention to the physical meaning. At the same time, similarity index fusion methods are springing up [5,6]. Recent years, with the development of deep learning, some deep features extraction methods have been proposed [7,8], the fusion of structure and attribute information has been attached importance again [9]. These methods have strong consistency. We divide link prediction method into single and combination methods, based on whether they use multidimension information, and whether they define the relation of multidimension information directly. For example, single methods, such as RA index [10], which defines the relation of common neighbors and degree of nodes directly; and classification based methods, index fusion methods, fusion of structure and attribute information methods belong to link prediction combination methods.

Most combination methods perform better than single methods, and are robust to many network types. However, what is the reason for this improved accuracy and robustness, and is there a theoretical limit for combination methods? This paper proposes the mathematic description of combination methods, and obtains the necessary and sufficient conditions for theoretical limit. The limit theorem also has important practical application value. It reveals the ultimate goal of combination methods that is to estimate probability density functions of existing links and nonexistent links. Thus, an appropriate form of the transformation function could be selected from the complete set. Based on limit theorem, a new combination method, theoretical limit fusion (TLF)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2017 doi:10.20944/preprints201710.0163.v1

method, is proposed. We use Parzen kernel method [11] of destiny estimation in the TLF method. Simulations and empirical studies have shown that TLF method can achieve higher prediction accuracy.

Section 2 introduces a mathematical description for the theoretical limit of combination methods and evaluation metrics for link prediction. Section 3 proposes and proves necessary and sufficient conditions for the theoretical limit of combination methods. Section 4 proposes a fusion link prediction method based on limit theorem (TLF method). Section 5 provides simulation examples for limit theorem and proposed TLF method with other combination methods, and gives comparison experiments in real networks. Section 6 and 7 discuss some results and conclude the paper.

2. Problem Description and Evaluation Metrics

2.1. Problem description

Given a network $G(V,E)$ at time t, where $V=\{v_1,v_2,\cdots,v_N\}$ is the set of nodes and $E=\{e_1,e_2,\cdots,e_M\}$ is the set of links. The observed links, E, are randomly divided into training, E^T, and probe, E^P, sets, where $E=E^T\bigcup E^P$ and $E^T\bigcap E^P=\varnothing$. Link prediction aims to predict missing links at current network or new links for a future time $t'(t'>t)$ [2]. Link prediction combination methods fuse several similarity indices and obtain a synthetic index and can be described in mathematic as follows. Let $\mathbf{X}=(X_1,X_2,\cdots,X_n)^T$ be the scores of existing links as given by n structural similarity indices, and follow probability density function (pdf) $f(\boldsymbol{x})=f(x_1,x_2,\cdots,x_n)$. Let $\mathbf{Y}=(Y_1,Y_2,\cdots,Y_n)^T$ be the scores of nonexistent links as n structural similarity indices, and follow $g(\boldsymbol{x})=g(x_1,x_2,\cdots,x_n)$. We need to find the transformation function, $l(\boldsymbol{x})$, and obtain the synthetic score, $X=l(\mathbf{X})$, $Y=l(\mathbf{Y})$ that maximizes evaluation metrics. Figure 1 is the diagram of combination methods.

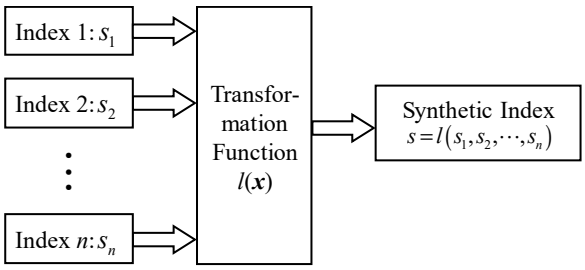

Figure 1. Combination methods

2.2. Evaluation metrics

Let the synthetic score $X=l(\mathbf{X})$ follow pdf $f_X(x)$, and $Y=l(\mathbf{Y})$ follow $g_Y(x)$. X and Y are independent. We have the following metrics.

2.2.1. Area under the receiver operation characteristics curve (AUC)

A receiver operating characteristics (ROC) curve is a two-dimensional depiction of classifier performance [12]. In the field of link prediction, the ROC curve abscissa represents the probability of nonexistent links i.e., the false positive rate (FPR), when the link prediction score is greater than some threshold, μ, and $\text{FPR}=\int_{\mu}^{\infty}g_Y(x)dx$. The ordinate represents the probability of missing links, i.e., the true positive rate (TPR), when score $>\mu$, and $\text{TPR}=\int_{\mu}^{\infty}f_X(x)dx$, TPR is equivalent to Recall. According to [13], AUC can be derived as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2017 doi:10.20944/preprints201710.0163.v1

$$
\begin{aligned}
P(X>Y) &= \iint_{X>Y} f_X(x)g_Y(y)dxdy \\
&= \frac{1}{2}\iint_{X>Y} f_X(x)g_Y(y)dxdy + \frac{1}{2}\left(1-\iint_{X\le Y} f_X(x)g_Y(y)dxdy\right) \\
&= \frac{1}{2}\iint \operatorname{sgn}(x-y) f_X(x)g_Y(y)dxdy + \frac{1}{2} \\
&= \frac{1}{2}\mathbb{E}\left[\operatorname{sgn}(X-Y)+1\right],
\end{aligned}
\tag{1}
$$

where

$$
\operatorname{sgn}(x) = \begin{cases} 1, & x>0 \\ 0, & x=0 \\ -1, & x<0 \end{cases}. \tag{2}
$$

In the real network, original data is randomly divided into training set and the probe set. Eq. (1) means that for n independent comparisons, if there are n' comparisons where the missing link returns a higher score and n'' comparisons where the missing and nonexistent links return the same score, we can obtain the algorithm expression of AUC:

$$
\mathrm{AUC} = \frac{n' + 0.5n''}{n}. \tag{3}
$$

2.2.2. Precision

Precision can be defined as the ratio of correct to (correct and error) prediction proportions when score > μ, i.e.,

$$
\begin{aligned}
\mathrm{Precision} &= \frac{P(\omega_1)\int_{\mu}^{+\infty} f_X(x)dx}{P(\omega_1)\int_{\mu}^{+\infty} f_X(x)dx + P(\omega_2)\int_{\mu}^{+\infty} g_Y(x)dx} \\
&= \frac{P(\omega_1)\mathrm{TPR}}{P(\omega_1)\mathrm{TPR} + P(\omega_2)\mathrm{FPR}}.
\end{aligned}
\tag{4}
$$

In the real network, if the top L links are predicted ones, with m links being right (i.e., there are m links in E^P), then

$$
\mathrm{Precision} = \frac{m}{L}. \tag{5}
$$

Owing to the imbalance of positive and negative samples, link prediction usually uses AUC metric. In application, high Precision means target links are accurate, and these links can be used directly. AUC and Precision are two important metrics in link prediction, we will study the theoretical limit using the two metrics.

3. Theoretical Limit Theorem

Theorem. *Let* $\mathbf{X} = (X_1, X_2, \cdots, X_n)^T$ *and* $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)^T$ *be random vectors following the joint distributions* $f(\boldsymbol{x})$ *and* $g(\boldsymbol{x})$, *respectively, where* $m\{\boldsymbol{x}: f(\boldsymbol{x})/g(\boldsymbol{x}) = C, g(\boldsymbol{x}) \neq 0,\ \forall C \in \mathbb{R}\} = 0$. (*m represents the measure of a set.*) *Then the following conditions are equivalent.*

(a) *A monotonically increasing function* $r(x)$ *exists, such that* $l(\boldsymbol{x}) = r[f(\boldsymbol{x})/g(\boldsymbol{x})], g(\boldsymbol{x}) \neq 0$, *a.e.* $\boldsymbol{x} \in \mathbb{R}^n$.

(b) *Transformation function* $l(\boldsymbol{x})$ *produces maximum AUC.*

If we add a condition in Theorem that prior probability of existing and nonexistent links be $P(\omega_1)$ *and* $P(\omega_2)$, *respectively. Then the following conditions are equivalent to* (a) *and* (b):

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2017 doi:10.20944/preprints201710.0163.v1

(c) *for any* α, *there exists the corresponding threshold* μ_l *for transformation* $l(\boldsymbol{x})$, *and satisfies* $\alpha = P(\omega_1)\int_{\mu_l}^{+\infty} f_X(x)dx + P(\omega_2)\int_{\mu_l}^{+\infty} g_Y(x)dx$, *such that transformation function* $l(\boldsymbol{x})$ *produces maximum Precision.*

Proof. (a) $\Rightarrow$ (b):

From the equivalent definition, AUC maximum is the maximum area under the ROC curve. For any FPR, if the TPRs corresponding to the ROC curve reach maximum, then the AUC reaches the maximum, i.e.,

$$\mathrm{FPR} = \int_{\mu}^{\infty} g_Y(x)dx = \int_{E(l(\boldsymbol{x})>\mu)} g(\boldsymbol{x})d\boldsymbol{x}, \tag{6}$$

$$\mathrm{TPR} = \int_{\mu}^{+\infty} f_X(x)dx = \int_{E(l(\boldsymbol{x})>\mu)} f(\boldsymbol{x})d\boldsymbol{x}, \tag{7}$$

where $E\left(l(\boldsymbol{x})>\mu\right)$ is a set $\left\{\boldsymbol{x}\in\mathbb{R}^n:\ \mu\in\mathbb{R},\ l(\boldsymbol{x})>\mu\right\}$, and $m\left\{\boldsymbol{x}:\ l(\boldsymbol{x})=C,\ \forall C\in\mathbb{R}\right\}=0$.

We use Lagrange's undetermined multipliers to solve this problem. For any specified FPR (denoted as FPR_0), the TPR corresponding to the ROC curve reaches maximum is equivalent as φ reaches maximum,

$$\begin{aligned}\varphi &= \int_{E(l(\boldsymbol{x})>\mu)} f(\boldsymbol{x})d\boldsymbol{x} + \lambda\left[\mathrm{FPR}_0 - \int_{E(l(\boldsymbol{x})>\mu)} g(\boldsymbol{x})d\boldsymbol{x}\right] \\ &= \lambda\mathrm{FPR}_0 + \int_{E(l(\boldsymbol{x})>\mu)}\left[f(\boldsymbol{x})-\lambda g(\boldsymbol{x})\right]d\boldsymbol{x}.\end{aligned} \tag{8}$$

Function φ will be maximized if we choose set $E(l(\boldsymbol{x})>\mu)$ such that the integrand is positive, i.e., if

$$f(\boldsymbol{x})-\lambda g(\boldsymbol{x})>0, \tag{9}$$

then $\boldsymbol{x}\in E(l(\boldsymbol{x})>\mu)$. Which means, no matter what is λ, if we select the set of $\boldsymbol{x}$ which makes the integrand $f(\boldsymbol{x})-\lambda g(\boldsymbol{x})$ always be positive, the function φ will reach maximum; if the set contains $\boldsymbol{x}$ that makes the integrand be negative, function φ will decrease. Let $l(\boldsymbol{x})=f(\boldsymbol{x})/g(\boldsymbol{x})$ and $\mu=\lambda$, and the set, $E(l(\boldsymbol{x})>\mu)$, equals to $E(f(\boldsymbol{x})/g(\boldsymbol{x})>\lambda)$, which satisfies (8), i.e.

$$\varphi = \lambda\mathrm{FPR}_0 + \int_{E(f(\boldsymbol{x})/g(\boldsymbol{x})>\lambda)}\left[f(\boldsymbol{x})-\lambda g(\boldsymbol{x})\right]d\boldsymbol{x}. \tag{10}$$

Thus, for any FPR, the TPR corresponding to the ROC curve reaches the maximum, so the AUC reaches the maximum when **X** and **Y** are transformed by $l(\boldsymbol{x})=f(\boldsymbol{x})/g(\boldsymbol{x})$.

Let $r(x)$ be a monotonically increasing function; and $h(x)$ be the inverse function of $r(x)$. If $h'(x)=1/r'(x)$, then $h(x)$ and $r(x)$ have the same monotonicity, and both are increasing functions. Thus, $|h'(x)|=h'(x)$. The pdf of $X_2=r(X_1)$ is $f_{X_2}(x)=f_{X_1}[h(x)]h'(x)$, and the pdf of $Y_2=r(Y_1)$ is $g_{Y_2}(x)=g_{Y_1}[h(x)]h'(x)$. Thus,

$$\begin{aligned}\mathrm{AUC} = P(X_2>Y_2) &= \int_{-\infty}^{+\infty} f_{X_2}(x)\int_{-\infty}^{x} g_{Y_2}(y)dydx \\ &= \int_{-\infty}^{+\infty} f_{X_1}\left(h(x)\right)h'(x)\int_{-\infty}^{x} g_{Y_1}\left(h(y)\right)h'(y)dydx \\ &= \int_{-\infty}^{+\infty} f_{X_1}(x)\int_{-\infty}^{x} g_{Y_1}(y)dydx \\ &= P(X_1>Y_1).\end{aligned} \tag{11}$$

We have proved (a) $\Rightarrow$ (b).

(b) $\Rightarrow$ (a): If $l_2(\boldsymbol{x})\neq r[l(\boldsymbol{x})]$, where $r(x)$ is increasing function, there exists $l_2(\boldsymbol{x})$ such that $\mathbf{X},\mathbf{Y}$ transforming from $l_2(\boldsymbol{x})$ can also produce maximum AUC, and then the corresponding ROC curves are the same. Otherwise, if ROC curves are different, except the same part, for any FPR, there is at least a ROC curve which doesn't reach maximum TPR, and contradict with maximum AUC. Since $m\{\boldsymbol{x}: f(\boldsymbol{x})/g(\boldsymbol{x})=C,\ g(\boldsymbol{x})\neq 0,\ \forall C\in\mathbb{R}\}=0$ and the ROC curve is the same for any point $(\mathrm{FPR},\mathrm{TPR})$ on the two ROC curves, thus,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2017 doi:10.20944/preprints201710.0163.v1

i. For any $\mathrm{FPR}\in[0,1]$, and any μ_{FPR}, there exist $\mu_{2\mathrm{FPR}}$, such that $E(l(\boldsymbol{x})>\mu_{\mathrm{FPR}})=E(l_2(\boldsymbol{x})>\mu_{2\mathrm{FPR}})$ for a.e. $\boldsymbol{x}\in\mathbb{R}^n$;

ii. For any $\mu^*_{\mathrm{FPR}}>\mu_{\mathrm{FPR}}$, if $E(l(\boldsymbol{x})>\mu^*_{\mathrm{FPR}})=E(l_2(\boldsymbol{x})>\mu^*_{2\mathrm{FPR}})$ and $E(l(\boldsymbol{x})>\mu_{\mathrm{FPR}})=E(l_2(\boldsymbol{x})>\mu_{2\mathrm{FPR}})$, then $\mu^*_{2\mathrm{FPR}}>\mu_{2\mathrm{FPR}}$.

Let $y_1=l(\boldsymbol{x})$, then a set of y_1 exist with nonzero measure, such that $l_2(\boldsymbol{x})\neq r[l(\boldsymbol{x})]$, i.e., $m\{y_1:l_2(\boldsymbol{x})\neq r[l(\boldsymbol{x})]\}\neq 0$. Let $\sigma=\{y_1:l_2(\boldsymbol{x})\neq r[l(\boldsymbol{x})]\}$. If $y_1\in\sigma$, $l_2(\boldsymbol{x}),l_1(\boldsymbol{x})$ satisfies function relation $l_2(\boldsymbol{x})=s[l(\boldsymbol{x})]$, but $s(x)$ is not increasing, then for any $\mu_1\in\sigma$, condition (ii) does not hold. If $y_1\in\sigma$, $l_2(\boldsymbol{x})$ and $l(\boldsymbol{x})$ are not functionally related, then neither condition (i) or (ii) hold. Thus (b) $\Rightarrow$ (a) is established.

(c) $\Leftrightarrow$ (b) : Let $k=\mathrm{TPR}/\mathrm{FPR}$ be the slope of the secant for any point on the ROC curve to the origin, then $\mathrm{Precision}=k/(k+\lambda)$, $\lambda=P(\omega_2)/P(\omega_1)$. For any α, that $l(\boldsymbol{x})$ produces maximum Precision is equivalent that k reaches maximum. And equivalent that for any α, $\alpha=P(\omega_1)\int_{\mu_l}^{+\infty}f_X(x)dx+P(\omega_2)\int_{\mu_l}^{+\infty}g_Y(x)dx$, $\mathrm{TPR}=\int_{\mu_l}^{+\infty}f_X(x)dx$ is maximum. Since this condition is established for any α, then it is equivalent that for any $\mathrm{FPR}\in[0,1]$, the corresponding TPR reaches maximum, and equivalent to $l(\boldsymbol{x})$ produces maximum AUC.

4. A fusion link prediction method based on limit theorem

The limit theorem of combination method shows that when selecting transformation function as $l(\boldsymbol{x})=f(\boldsymbol{x})/g(\boldsymbol{x})$ or its monotone increasing transformation, the AUC and Precision of synthetic score reaches the maximum. In the real network, because $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are unknown, the pdfs need to be estimated from multidimensional data. Let the estimated pdfs be $\hat{f}(\boldsymbol{x})$ and $\hat{g}(\boldsymbol{x})$. On the basis of limit theorem, we define the transformation function as the ratio of estimated pdfs, i.e.,

$$\hat{l}(\boldsymbol{x})=\hat{f}(\boldsymbol{x})/\hat{g}(\boldsymbol{x}). \tag{12}$$

Then we obtained the synthetic score, $s=\hat{l}(\boldsymbol{x})$, and used for link prediction. This method is called theoretical limit fusion (TLF) method.

Before evaluating $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$, the input link prediction scores need to be normalized,

$$s_k^*(i,j)=\frac{0.5N^2\cdot s_k(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}s_k(i,j)},k=1,2,\cdots,n. \tag{13}$$

$s_k(i,j)$ represents the k-th similarity score for node pair i,j. N is the dimension of adjacent matrix, and n is the number of similarity indices.

The limit theorem of combination method transformed the link prediction indices fusion problem into the estimation of pdfs. Statistical methods for estimating density functions can be applied to this problem, directly. Parzen kernel method [11] of destiny estimation is used in this paper. The multivariate kernel density estimate defined as:

$$\hat{f}(\boldsymbol{x})=\frac{1}{nh^d}\sum_{i=1}^{n}K\left[\frac{1}{h}(\boldsymbol{x}-\boldsymbol{x}_i)\right], \tag{14}$$

where h is the window width, and $K(\boldsymbol{x})$ is a multivariate kernel defined for d-dimensional $\boldsymbol{x}$, such that

$$\int_{\mathbb{R}^d}K(\boldsymbol{x})d\boldsymbol{x}=1. \tag{15}$$

A form of the pdf estimate commonly used is Gauss kernel,

$$K(\boldsymbol{x})=\frac{1}{(2\pi)^{d/2}}\exp\left(\frac{-\boldsymbol{x}^T\boldsymbol{x}}{2}\right). \tag{16}$$

In summary, the steps of TLF are listed as Table 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2017 doi:10.20944/preprints201710.0163.v1

Table 1. The steps of TLF method.

Step 1	Divide the network into training set, E^T, and probe set, E^P;
Step 2	According to Eq. (13), normalize these similarity indices, then we distinguish existing links and nonexistent links in the training set;
Step 3	Based on Eq. (14), estimate the pdfs of existing links and nonexistent links, and we obtain the estimated pdfs as $\hat{f}(\boldsymbol{x})$ and $\hat{g}(\boldsymbol{x})$;
Step 4	Obtain the synthetic score of n structure similarity indices according to Eq. (12);
Step 5	Calculate the accuracy such as AUC metric or Precision metric on the probe set.

5. Simulation and experiment

5.1. Simulation examples

Four types of structural similarity indices were simulated to evaluate node pairs with and without links. The pdfs of the structural similarity indices are also provided. We construct 3 groups of known distributions for the similarity indices pdfs. One thousand samples extracted from 10000 existing links and 100000 samples of nonexistent links were generated following the appropriate pdfs. The 1000 samples serve as probe set; the 100000 samples with 1000 probe links serve as unknown links for training; and the $(10000-1000)$ samples serve as train set of existing links. Each sample had 4 dimensions to simulate 4 similarity scores. We first compute AUC and Precision for each dimension, then use proposed TLF method to obtain the synthetic score and calculate the AUC and Precision, compared with other combination methods such as Naïve Bayes and logistic regression. Finally, we calculate AUC and Precision using the theoretical limit theorem and compare with the above methods.

Let random vectors $\mathbf{X}=(X_1,X_2,X_3,X_4)^T$ and $\mathbf{Y}=(Y_1,Y_2,Y_3,Y_4)^T$ be the scores of existing and nonexistent links, which follow $f(\boldsymbol{x})=f(x_1,x_2,\cdots,x_n)$ and $g(\boldsymbol{x})=g(x_1,x_2,\cdots,x_n)$ pdfs, respectively.

Let $f(\boldsymbol{x}),g(\boldsymbol{x})$ are 4-dimensional normal distributions,

$$f(\boldsymbol{x})=\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \tag{17}$$

where $\mathrm{diag}(\boldsymbol{\Sigma})\mathbf{1}=\left(\sigma_1^2,\sigma_2^2,\sigma_3^2,\sigma_4^2\right)^T$, and $\boldsymbol{\Sigma}_{ij}=\boldsymbol{r}_{ij}\sigma_i\sigma_j$.

The parameter sets for the 2 groups of simulation examples are as follows.

Group 1: $\Theta_{1f}=\{\boldsymbol{\mu}_{1f},\boldsymbol{\Sigma}_{1f}\}$, and $\Theta_{1g}=\{\boldsymbol{\mu}_{1g},\boldsymbol{\Sigma}_{1g}\}$;

Group 2: $\Theta_{2f}=\{\boldsymbol{\mu}_{2f},\boldsymbol{\Sigma}_{2f}\}$ and $\Theta_{2g}=\{\boldsymbol{\mu}_{2g},\boldsymbol{\Sigma}_{2g}\}$.

In each group, $\boldsymbol{\mu}_{1f}=(1,2,1.7,2.1)^T$, $\boldsymbol{\mu}_{1g}=(1.3,2.5,2.1,2.8)^T$, $\boldsymbol{\mu}_{2f}=(1,2,1.7,2.1)^T$, $\boldsymbol{\mu}_{2g}=(1.5,3.5,2.8,3)^T$, $\mathrm{diag}\left(\boldsymbol{\Sigma}_{1f}\right)\mathbf{1}=\left(1.5^2,2.2^2,3^2,2.5^2\right)^T$, $\mathrm{diag}\left(\boldsymbol{\Sigma}_{1g}\right)\mathbf{1}=\left(2^2,2.2^2,3^2,2.5^2\right)^T$, $\mathrm{diag}\left(\boldsymbol{\Sigma}_{2f}\right)\mathbf{1}=\left(1.5^2,2.2^2,3^2,2.5^2\right)^T$, $\mathrm{diag}\left(\boldsymbol{\Sigma}_{2g}\right)\mathbf{1}=\left(2.5^2,3.5^2,4^2,2.5^2\right)^T$,

$$\boldsymbol{r}_{1f}=\boldsymbol{r}_{1g}=\begin{bmatrix}1 & 0.8 & 0.76 & 0.56\\ 0.8 & 1 & 0.85 & 0.74\\ 0.76 & 0.85 & 1 & 0.93\\ 0.56 & 0.74 & 0.93 & 1\end{bmatrix},$$

and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2017 doi:10.20944/preprints201710.0163.v1

$$\boldsymbol{r}_{2f} = \boldsymbol{r}_{2g} = \begin{bmatrix} 1 & 0.62 & 0.45 & 0.34 \\ 0.62 & 1 & 0.28 & 0.47 \\ 0.45 & 0.28 & 1 & 0.65 \\ 0.34 & 0.47 & 0.65 & 1 \end{bmatrix}.$$

The window width h of TLF method in the group 1 and 2 is $h = 0.1$.

Group3. Let

$$f_3(\boldsymbol{x}) = x_1x_2x_3x_4 + x_1x_4 + x_3\exp(x_1)\log(x_2)$$
$$(0 \le x_1 \le 3, 1 \le x_2 \le 3, 3 \le x_3 \le 5, 2 \le x_4 \le 3.5)' \quad (18)$$

and

$$g_3(\boldsymbol{x}) = x_1x_2x_3x_4 + x_3\exp(x_1)\log(x_2)$$
$$(0 \le x_1 \le 4, 1 \le x_2 \le 3, 3 \le x_3 \le 5, 2.5 \le x_4 \le 5)' \quad (19)$$

We ignore the constant that makes the integral of $f(\boldsymbol{x}), g(\boldsymbol{x})$ equal to 1. The simulation results of group 3 are shown as Table. 2.

Table 2. Simulation results of group 1 and group 2.

Parameters	Accuracy	Dim1	Dim2	Dim3	Dim4	NB	LR	TLF	Theoretical Limit	Transform by increasing function
Group 1	AUC	0.554	0.566	0.547	0.585	0.610	0.668	**0.691**	0.738	0.738
	Precision	0.047	0.015	0.014	0.027	0.038	0.020	**0.097**	0.120	0.120
Group 2	AUC	0.569	0.660	0.604	0.622	0.765	0.676	**0.786**	0.792	0.792
	Precision	0.114	0.140	0.081	0.038	0.153	0.051	**0.212**	0.241	0.241

The simulation results in Table 2 and Table 3 show us that we can calculate the theoretical limit of combination method based on theorem 1, and the limit AUC and Precision are highest among all listed methods, though we cannot list all possible conditions. Results also show that TLF method can fuse the information effectively, and obtain the optimum accuracy. We also verify that the transformation of monotonically increasing function does not change the theoretical limit. Theorem 1 provides a platform that can compare each combination method by constructing some distributions, and direct an effect combination method TLF.

Table 3. Simulation results of group 3.

Accuracy	Dim1	Dim2	Dim3	Dim4	NB	LR	TLF	Theoretical Limit	Transform by increasing function
AUC	0.770	0.505	0.488	0.878	0.938	0.923	**0.950**	0.956	0.956
Precision	0.567	0.007	0.007	0.654	0.711	0.100	**0.815**	0.858	0.858

The window width h of TLF method in the group 3 is $h = 0.1$.

5.2. Experiments in real networks

The significance of simulation is that the theoretical limit can be derived by theoretical calculation or numerical calculation, and all combination methods can be used to compare with it, finding shortcomings and gaps to design a more rational method. However, the simulation data is different from real network data. We use TLF method to fuse several similarity indices and test in real networks. The basic similarity indices we use are Common Neighbor index (CN) [14], Adamic-Adar index (AA) [15], Resource Allocation index (RA) and Preferential Attachment index (PA) [16,17]. These indices are local indices. Several global indices such as Katz index [18], Average Commute Time index (ACT) and Cosine Similarity Time index (Cos+) are served as comparisons [19,20]. The definitions of the above indices and their meanings are listed as Table 4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2017 doi:10.20944/preprints201710.0163.v1

Table 4. Definitions and descriptions of similarity indices.

Index	Equation	Description
CN	$s_{\mathrm{CN}}(i,j)=\lvert\Gamma(i)\cap\Gamma(j)\rvert$	$\Gamma(i)$ is the set of neighbors of node i. $\lvert\cdot\rvert$ represents cardinality of a set. CN index denotes the common neighbors between nodes i and j.
AA	$s_{\mathrm{AA}}(i,j)=\sum_{z\in\Gamma(i)\cap\Gamma(j)}\frac{1}{\log k_z}$	AA index weights the common neighbors by the reciprocal of the logarithm of each node's degree.
RA	$s_{\mathrm{RA}}(i,j)=\sum_{z\in\Gamma(i)\cap\Gamma(j)}\frac{1}{k_z}$	RA index weights the common neighbors by the reciprocal of each node's degree.
PA	$s_{\mathrm{PA}}(i,j)=k_i k_j$	PA index expresses preferential attachment by node's degree.
Katz	$s_{\mathrm{Katz}}(i,j)=\left[\lim_{n\to\infty}\sum_{m=1}^{n}(\alpha\mathbf{A})^m\right]$	**A** is adjacent matrix of network. Katz index considers all path between two nodes and gives more weights, α, to the shorter paths.
ACT	$s_{\mathrm{ACT}}(i,j)=\frac{1}{l_{ii}^{+}+l_{jj}^{+}-2l_{ij}^{+}}$	l_{xy}^{+} is the corresponding element in $\mathbf{L}^{+}$, and $\mathbf{L}^{+}$ denotes the pseudo-inverse of laplacian matrix.
Cos+	$s_{\mathrm{Cos+}}(i,j)=\frac{v_i^T v_j}{\lvert v_i\rvert\cdot\lvert v_j\rvert}=\frac{l_{ij}^{+}}{\sqrt{l_{ii}^{+}\cdot l_{jj}^{+}}}$	According to $\mathbf{L}^{+}$, Cos+ calculates cosine similarity of two vectors in matrix $\mathbf{L}^{+}$.

We use TLF method to fuse 4 local similarity indices, and compare with fusion method such as naïve Bayes and logistic regression and other global indices. Our experiments are performed on 8 different real networks. (1) FWEW [21], (2) FWFB [22], (3) PPI_Cell [23], (4) CKM-3 [24], (5) netscience (NS) [25], (6) Yeast [26], (7) PB [27], (8) email [28]. The basic topological features of 8 real networks are listed in table 5. Each original data is randomly divided into training set of 90% links, and the probe set of 10% links.

Table 6 and Table 7 show the comparisons between TLF method and other combination methods or global indices using AUC and Precision metrics. Each result is the average of 10 realizations.

Table 5. Basic topological features of 6 example networks. $|V|$ and $|E|$ are the total numbers of nodes and links, respectively. $\langle k\rangle$ represents the average degree of nodes in a network, and $\langle d\rangle$ represents the average distance between nodes in a network. C and r are the clustering coefficient and assortative coefficient respectively. H is the degree heterogeneity, defined as $H=\frac{\langle k^2\rangle}{\langle k\rangle^2}$.

Data	$\lvert V\rvert$	$\lvert E\rvert$	$\langle k\rangle$	$\langle d\rangle$	r	C	H
FWEW	69	880	25.51	1.636	-0.298	0.560	1.275
FWFB	128	2075	32.42	1.78	-0.112	0.335	1.24
PPI_Cell	127	237	3.732	4.450	0.035	0.455	1.649
CKM-3	246	423	3.439	4.240	0.102	0.356	1.335
Yeast	2375	11693	9.85	5.10	0.469	0.378	3.48
PB	1222	16717	27.36	2.74	-0.221	0.361	2.97
NS	1589	2742	3.451	1.333	0.462	0.889	2.011
Email	1133	5451	9.622	3.606	0.078	0.297	1.942

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2017 doi:10.20944/preprints201710.0163.v1

Table 6. Comparisons of the AUC value between TLF and other combination methods or global indices. In each network, the selected window width *h* is along with the AUC value.

Data	CN	AA	RA	PA	ACT	Cos+	Katz	NB	LR	TLF
FWEW	0.687	0.694	0.714	0.819	0.793	0.511	0.727	0.825	0.832	**0.876** (*h*=0.1)
FWFB	0.624	0.624	0.624	0.742	0.727	0.649	0.680	0.749	0.762	**0.781** (*h*=0.1)
PPI_Cell	0.736	0.745	0.740	0.699	0.779	0.783	0.822	0.753	0.679	**0.831** (*h*=0.3)
CKM-3	0.661	0.665	0.661	0.585	0.560	0.535	**0.928**	0.683	0.675	0.713 (*h*=0.15)
Yeast	0.918	0.918	0.915	0.869	0.903	0.958	0.962	0.925	0.934	**0.968** (*h*=0.2)
PB	0.922	0.928	0.928	0.906	0.890	0.932	0.934	0.931	0.936	**0.949** (*h*=0.3)
NS	0.994	0.994	0.995	0.709	0.558	0.507	0.996	0.998	**0.999**	**0.999** (*h*=0.2)
Email	0.849	0.852	0.851	0.817	0.801	0.889	0.908	0.865	0.870	**0.912** (*h*=0.15)

Table 7. Comparisons of the Precision value between TLF and other combination methods or global indices. In each network, the corresponding window width *h* is the same as Table 6.

Data	CN	AA	RA	PA	ACT	Cos+	Katz	NB	LR	TLF
FWEW	0.143	0.145	0.162	0.334	0.271	0.004	0.196	0.301	0.325	**0.543**
FWFB	0.071	0.072	0.083	0.240	0.184	0.029	0.148	0.249	0.283	**0.382**
PPI_Cell	0.052	0.048	0.073	0.012	0.045	0.061	0.058	0.072	0.068	**0.085**
CKM-3	0.051	0.059	0.062	0.011	0.001	0.003	0.061	0.060	0.062	**0.064**
Yeast	0.652	0.703	0.461	0.439	0.487	0.291	0.721	0.712	0.723	**0.785**
PB	0.381	0.320	0.212	0.100	0.129	0.298	0.381	0.411	0.395	**0.452**
NS	0.820	0.971	0.982	0.008	0.004	0.006	0.823	0.988	0.986	**0.991**
Email	0.202	0.253	0.214	0.039	0.031	0.086	0.231	0.263	0.289	**0.347**

The results show us that TLF method performs better than other fusion methods such as naïve Bayes and logistic regression, no matter what evaluation metric use. Almost all combination methods are better than 4 basic indices. From the limit theorem, combination methods are dependent with each dimension. The promotion of fusion index is restrict to each similarity index. Experiment results also exposed this problem: if the single similarity indices perform not well, the fusion index cannot significantly improve the accuracy. For example, in the CKM-3 network, though we use TLF method to fuse 4 basic similarity indices can improve the AUC obviously, it cannot be better than Katz index (0.928).

6. Discussion

Many combination methods try to find the nonlinear relation of every dimensions, and want to obtain a more reasonable fusion function to promote the prediction accuracy. For example, link prediction method based on the choquet fuzzy integral [5] uses fuzzy measures to measure the importance of each similarity index in the fusion process and the interaction between them. Logistic regression based index adopts logistic function to learn the relation of multiple structural features and obtain an adaptive link prediction method [29]. In fact, according to the limit theorem, the nonlinear relation is the ratio of two joint probability destiny functions or its monotone increasing transformation. The best fusion function is a measurement of difference between existing and nonexistent links, and it reflects the relativity of existing and nonexistent links. The essence of combination methods is trying to approximate the pdfs from many aspects. Limit theorem provides a unified interpretation for all combination methods. On the basis of theoretical limit theorem, the

proposed TLF method evaluates two pdfs directly, and it has a better fusion effect from results of simulation and experiment in real network.

5. Conclusions

This paper proposes mathematic description of link prediction combination methods and derives the limit theorem. Before the mathematic description we proposed, many combination methods have been put forward and widely used. However, all these methods are groping respectively without unified explanation. Limit theorem solved this problem and provided a guidance for link prediction method design. The TLF method based on limit theorem can achieve higher prediction accuracy.

Acknowledgments: We acknowledge professor Guo'en Hu for inspirations. This work was partially supported by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (No. 61521003), and National Natural Science Foundation of China (No. 61601513).

Author Contributions: Yiteng Wu and Hongtao Yu proposed mathematical description of combination method; Yiteng Wu proposed and proved the theoretical limit theorem; Yiteng Wu and Ruiyang Huang designed the experiments and analyzed the results. Yingle Li and Senjie Lin wrote part of code.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Seife, C. What are the limits of conventional computing, *Science,* **2005**, 309.5731:96, DOI: 10.1126/science.309.5731.96
2. Wang, Peng, et al. Link prediction in social networks: the state-of-the-art, *Science China Information Sciences*, **2015**, 58.1:1–38, DOI: arXiv:1411.5118.
3. Lü L, Pan L, Zhou T, et al. Toward link predictability of complex networks. *Proceedings of the National Academy of Sciences*, **2015**, 112(8): 2325–2330, DOI: 10.1073/pnas.1424644112.
4. Lü L, Zhou T. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, **2011**, 390(6): 1150–1170, DOI: https://doi.org/10.1016/j.physa.2010.11.027.
5. Yu H T, Wang S H, Ma Q. Link prediction algorithm based on the Choquet fuzzy integral. *Intelligent Data Analysis*, **2016**, 20(4): 809–824, DOI: 10.3233/IDA-160833.
6. He Y, Liu J N K, Hu Y, et al. OWA operator based link prediction ensemble for social network. *Expert Systems with Applications*, **2015**, 42(1): 21–50, DOI: 10.1016/j.eswa.2014.07.018.
7. Liao, Lizi, et al. Attributed Social Network Embedding, *Transactions on Knowledge and Data Engineering*, **2017**.
8. Grover, Aditya, and J. Leskovec. node2vec: Scalable Feature Learning for Networks. *KDD*, **2016**:855.
9. Li, Wenjie, W. Li, and W. Li. Predictive Network Representation Learning for Link Prediction. *International ACM SIGIR Conference on Research and Development in Information Retrieval* ACM, **2017**:969-972.
10. Ou, Q., et al. Power-law strength-degree correlation from resource-allocation dynamics on weighted networks, *Physical Review E Statistical Nonlinear & Soft Matter Physics*, **2007**, 75.1:021102, DOI: 10.1103/PhysRevE.75.021102.
11. Parzen, E. On estimation of a probability density function and mode, *Ann. Math. Statist.*, **1962**, 33, 1065–1076.
12. Fawcett T. An introduction to ROC analysis. *Pattern recognition letters*, **2006**, 27(8): 861–874, DOI: 10.1016/j.patrec.2005.10.010.
13. Hanley J A, McNeil B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **1982**, 143(1): 29–36, DOI: 10.1148/radiology.143.1.7063747.
14. François Lorrain, Harrison C. White. Structural equivalence of individuals in social networks. Social Networks, **1977**, 1(1):67–98, DOI: 10.1080/0022250X.1971.9989788.
15. Adamic L A, Adar E. Friends and neighbors on the web. Social networks, **2003**, 25(3): 211–230.
16. Zhou T, Lü L, Zhang Y C. Predicting missing links via local information. The European Physical Journal B-Condensed Matter and Complex Systems, **2009**, 71(4): 623–630, DOI: 10.1140/epjb/e2009-00335-8.

17. Barabasi A L, Albert R. Emergence of scaling in random networks. Science, **1999**, 286(5439):509–512, DOI: 10.1126/science.286.5439.509.
18. J. Coleman, E. Katz, and H. Menzel. The Diffusion of an Innovation Among Physicians Sociometry, **1957**, 20:253–270, DOI: 10.2307/2785979.
19. Klein d J, Randić M. Resistance distance. Journal of Mathematical Chemistry, **1993**, 12(1):81–95, DOI: 10.1007/BF01164627.
20. Fouss F, Pirotte A, Renders J, et al. Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation. IEEE Transactions on Knowledge & Data Engineering, **2007**, 19(3):355-369, DOI: 10.1109/TKDE.2007.46.
21. R. E. Ulanowicz, D. L. DeAngelis, US Geological Survey Program on the South Florida Ecosystem, **2005**, 114.
22. Ulanowicz R E, Bondavalli C, Egnotovich M S. Network Analysis of Trophic Dynamics in South Florida Ecosystem, FY 97: The Florida Bay Ecosystem. Technical report, CBL, **1998**: 98–123.
23. E.D. Kolaczyk, Statistical Analysis of Network Data: Methods and Models, Springer, New York, **2009,** DOI: 10.1111/j.1751-5823.2010.00109_2.x · Source: RePEc.
24. Coleman J, Katz E, Menzel H. The Diffusion of an Innovation among Physicians 1. Social Networks, **1977**, 20(4):107-124, DOI: 10.2307/2785979.
25. Newman M E J. Finding community structure in networks using the eigenvectors of matrices. Physical Review E Statistical Nonlinear & Soft Matter Physics, **2006**, 74(3 Pt 2):036104, DOI: 10.1103/PhysRevE.74.036104.
26. Von Mering C, Krause R, Snel B, et al. Comparative assessment of large-scale data sets of protein protein interactions. Nature, **2002**, 417: 399–403, DOI: 10.1038/nature750.
27. Adamic L A, Glance N. The political blogosphere and the 2004 U.S. election: divided they blog, International Workshop on Link Discovery. ACM, **2005**:36-43, DOI: 10.1145/1134271.1134277.
28. R. Michalski, S. Palus, P. Kazienko, Matching Organizational Structure and Social Network Extracted from Email Communication, *Business Information Systems*, Springer, **2011**, pp. 197-206, DOI: 10.1007/978-3-642-21863-7_17.
29. Ma C, Bao Z K, Zhang H F. Improving link prediction in complex networks by adaptively exploiting multiple structural features of networks. Physics Letters A, **2017**.