

基于三元组结构的有向网链路预测方法

常圣, 马宏, 刘树新

(国家数字交换系统工程技术研究中心, 河南 郑州 450002)

摘要: 当前链路预测的研究主要集中在无向网络, 然而现实世界中存在大量的有向网络, 忽略链路的方向会缺失一些重要信息甚至使预测失去意义, 而直接将无向网络的预测方法应用于有向网络又存在预测精度降低的问题。为此, 提出了一个基于三元组的有向网络链路预测算法, 该算法针对有向网络和无向网络三元组结构的不同, 应用势理论对三元组进行筛选, 通过统计分析不同三元组闭合的可能性, 以网络整体三元组闭合指数作为权重计算节点间的相似性。在9个真实数据集上的实验表明, 所提方法比基准方法的预测精度提高了4.3%。

关键词: 链路预测; 有向网络; 三元组

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2019049

New method for link prediction in directed networks based on triad patterns

CHANG Sheng, MA Hong, LIU Shuxin

National Digital Switching System Engineering & Technological R & D Center, Zhengzhou 450002, China

Abstract: Almost all current studies on link prediction problem focus on undirected networks. Unfortunately, many complex networks in the real world are directed. Ignoring the direction of a link will overlook some important information or even make the prediction meaningless. Directly applying the methods for undirected networks to directed networks will reduce the accuracy of prediction. A new method for link prediction in directed networks based on triad patterns was proposed. The proposed metric compare the difference of triad structures between undirected and directed networks and use potential theory to filter the triad patterns. By statistics of triad closeness in various networks, new method calculate the similarity between nodes using the triad closeness index of a network as the weight for different triad patterns. Experiments on nine real networks show that accuracy of proposed method is 4.3% better than benchmark methods.

Key words: link prediction, directed networks, triad patterns

收稿日期: 2018-12-11; 修回日期: 2019-05-21

通信作者: 常圣, 724986365@qq.com

基金项目: 国家自然科学基金资助项目 (No.61803384)

Foundation Item: The National Natural Science Foundation of China (No.61803384)

论文引用格式: 常圣, 马宏, 刘树新. 基于三元组结构的有向网链路预测方法[J]. 网络与信息安全学报, 2019, 5(5): 39-47.
CHANG S, MA H, LIU S X. New method for link prediction in directed networks based on triad patterns [J]. Chinese Journal of Network and Information Security, 2019,5(5): 39-47.

1 引言

随着社交网络的兴起,链路预测受到越来越多的关注。链路预测是指通过已知的网络信息,预测网络中两个节点之间存在连边的可能性^[1]。这种预测既包含对未知连边(尚未观测到的连边)的预测,也包含对未来连边(尚未产生的连边)的预测。网络信息包括节点的属性和网络结构。虽然考虑节点的属性可以提高预测的准确性,但多数情况下节点的属性获取较为困难^[2],并伴有噪声,引入节点属性还会增加计算的复杂度。近年来,基于网络结构的方法受到学者越来越多的青睐。基于网络结构的方法大体可以分为两类^[2]:基于相似性的方法,概率和统计方法。概率和统计方法主要包括层次结构模型^[3]和随机分块模型^[4]等,此类方法虽然能够取得不错的预测效果,但计算复杂度较高,难以应用于规模较大的网络。与之相比,基于相似性的方法计算复杂度较低,且预测精度较好。共同邻居(CN, common neighbor)指标是最简单的相似性指标,其核心思想认为如果两个节点共同邻居越多,则它们之间存在连边的可能性越大。受到共同邻居的启发,加入对节点度的考虑,产生了很多相似性指标,包括 Adamic/Adar(AA)^[5]、Salton(SA)^[6]、Jaccard(JA)^[7]和资源分配指标(RA)^[8]等。此外,具有代表性的相似性指标还有 Katz^[9]指标和有重启的随机游走(RWR)^[10]等。

当前链路预测的研究工作对于无向网络的关注远远高于有向网络。然而现实世界中大部分网络是有向的,如食物链网络中的捕食关系,如果忽略连边的方向,可能会缺失重要的信息甚至失去链路预测的意义。在有向网络中,连边的方向让链路预测变得更加困难。文献[11-12]曾把部分相似性指标(CN、RA、AA等)直接应用于有向网络,这种方法可能会造成信息缺

失,因为不同方向的连边具有不同的含义,不同的形成机理在链路预测的过程中对于相似度的影响也是不同的。Narayanan等^[11]将局部随机游走推广到了有向网络,并取得了较好的预测效果,Lichtenwalter等^[13]基于随机游走提出了 PropFlow 方法,但是这两种方法相比局部性方法复杂度较高,难以应用于大型复杂网络。张千明等^[14]在 2013 年提出了势理论,并筛选出了具有较高预测精度的双风扇子图(bi-fan)结构。近几年来,基于三元组结构的方法受到越来越多的关注。文献[15]统计了 13 种三元组的频次,以此来计算节点间的相似度,如果某个类型的三元组出现越多,那么这种三元组对于相似度的贡献就越高。虽然该方法取得了不错的预测精度,但其精度是结合分类器得到的,指标本身对预测精度的贡献作者并未给出。文献[16]分别计算 9 种非闭合三元组结构的相似性,但是没有区分不同类型三元组的作用,而是直接将结果构成一个 9 维特征向量作为监督学习的输入。总体来看,当前对有向网络还缺乏深入的研究,多数方法是对无向网络算法的直接应用,或者结合机器学习方法来提高预测精度,缺乏对于网络本身内部机制的挖掘,因此,如何量化不同方向连边和不同结构的作用,从而设计一个有效的有向网络链路预测方法具有十分重要的意义。

基于上述情况,本文提出了一种基于三元组结构的有向网络链路预测方法。该方法根据有向网络和无向网络中三元组结构的差异、应用势理论,从 9 个非闭合三元组中筛选出 4 个进行相似度计算,通过统计分析不同三元组闭合的可能性,以网络整体三元组闭合指数为权重计算节点间的相似性。通过在 9 个不同性质的真实数据集上进行实验,并与基准方法进行比较,证明改进后的方法预测精度更高。

2 基于三元组的有向网链路预测方法

大多数基于相似性的链路预测算法是以三元组作为分析单位的。无向网络中三元组的节点关系非常简单, 需要考虑的未闭合结构只有图1中的一种情况, 所形成的闭合结构也仅有一种。

在有向网络中, 连边方向的出现使三元组结构变得复杂。有共同邻居且未闭合的三元组有9种, 形成新连边的情况更多。与无向网络相比, 有向网络中三元组的网络结构发生了较大变化。如果直接将无向网络的预测算法应用到有向网络, 那么无论使用图1中哪一种结构进行预测都会缺失大量的结构信息, 也会混淆出度和入度等概念, 显然这并不符合有向网的连边机理和演化规律。由此产生两个问题: ① 使用哪种结构进行预测能够获得较高的预测精度; ② 如何对每种结构进行量化。

2.1 三元组结构的筛选

张千明等^[14]在2013年提出了势理论, 该理论是对复杂网络中节点等级的一种描述。对于节点 x 和 y , 如果存在 x 指向 y 的连边, 则 x 的势能比 y 高一个单位。如果一个子图中每个节点的势能都能确定, 则称这个子图为可定义势的。显然, 含有互惠边的子图都是不可定义势的。图2展示了一些不可定义势和可定义势子图的例子。节点中的数字代表节点的势。

势理论认为, 若一条连边的出现能够产生更

多的可定义势子图, 那么它出现的可能性越大^[14]。在真实网络的实验也表明, 使用可定义势子图构造的预测器具有更高的预测精度。本文对此进行了延伸, 直接使用势理论对非闭合三元组(S1-S9)进行筛选: 9种三元组结构中S1-S4是可定义势的。同时考虑到S1-S4的计算复杂度比S5-S9低, 本文使用S1-S4结构进行相似度计算。

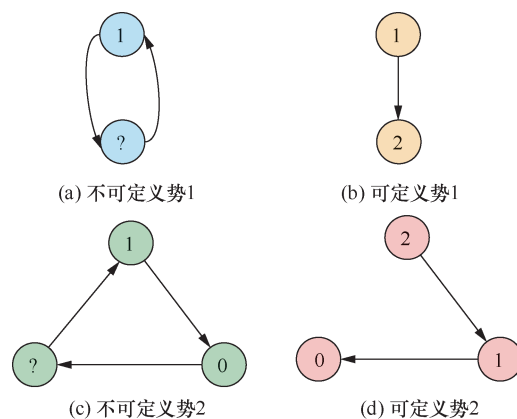


图2 不可定义势和可定义势子图示例

2.2 三元组结构权重的量化

现实网络中, 不同复杂网络之间结构和演化规律可能有很大不同, 如社交网络和食物链网络, 不同的三元组形成机理不同, 闭合的概率也不一定相同, 显然赋予不同三元组同样的权重是欠妥的。此外, 如果赋予不同结构一个固定的权重, 可能导致在不同性质网络中预测精度差异较大。如果想要同时提高预测精度和指标的适用范围, 可以将网络的某些自身特性作为预测指标的一部分。基于以上考虑, 本文将一个网络中某种三元组(S1-S9)的闭合指数(TCI, triad closeness index)

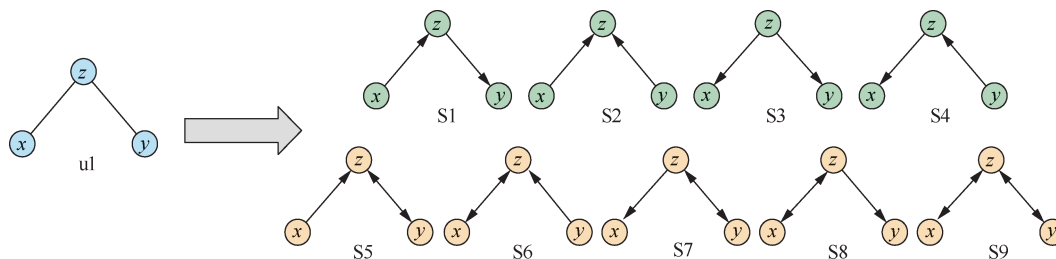


图1 无向图与有向图中未闭合三元组结构差异

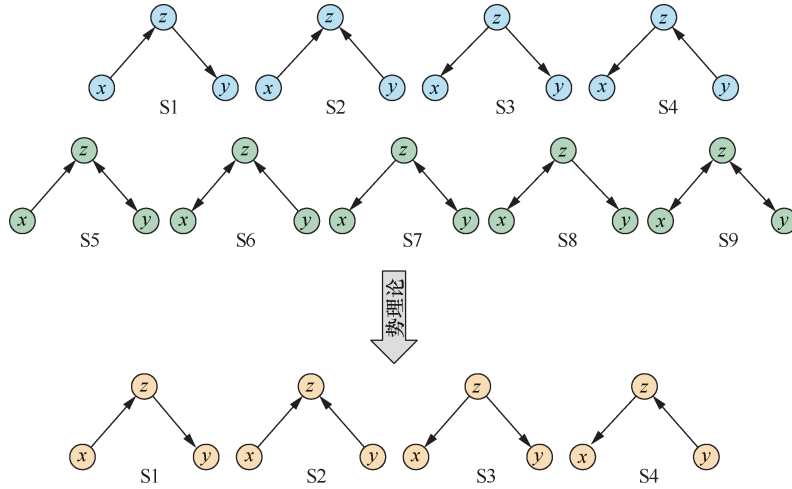


图 3 三元组结构的筛选

定义如下。

$$TCI_{xyz} = \frac{\text{countClose}(\text{pattern}(x, y, z))}{\text{count}(\text{pattern}(x, y, z))} \quad (1)$$

其中, $\text{pattern}()$ 获得节点 x 、 y 和 z 构成的三元组结构, $\text{count}()$ 统计某种三元组结构在整个网络中出现的频次, $\text{countClose}()$ 统计三元组对应闭合结构的出现频次。需要注意的是, 本文提到的闭合三元组仅指单方向的闭合。如图 4 所示, 考虑节点 x 和 y 及其共同邻居节点 z 所构成的三元组, 如果存在由节点 x 指向节点 y 的连边, 无论是否存在由节点 y 指向 x 的连边, 均称此三元组为闭合三元组, 反之则为未闭合三元组。这样定义的好处在于, 每次计算时只考虑一个方向的连边, 简单清晰, 且在进行矩阵运算时, 会遍历每个节点, “自动” 计算另一个方向连边的预测值。

2.3 预测方法

基于 2.1 节和 2.2 节的考虑, 每个 S1-S4 结构中共同邻居节点对于相似度的贡献为

$$w(x, y, z) = TCI_{xyz} \cdot f(z) \quad (2)$$

TCI_{xyz} 是 2.2 节中定义的三元组闭合指数, $f(z)$ 是节点贡献函数。在有向网络中, 节点的度数有出度和入度的区分。对于不同的三元组, 共同邻居节点的出度和入度对于新连边的影响程度是不同的。例如, S2 结构三元组的出现频次仅与

节点 z 的入度有关, 而 S3 结构则仅和 z 的出度相关。参照 RA 指标, 度数越大, 其共同邻居节点对相似度的贡献越小。基于以上考虑, $f(z)$ 等于

$$f(z) = \begin{cases} \frac{1}{|\Gamma_{\text{out}}(z)|}, & z \in S_2 \\ \frac{1}{|\Gamma_{\text{in}}(z)|}, & z \in S_3 \\ \frac{2}{|\Gamma_{\text{all}}(z)|}, & z \in S_1, S_4 \end{cases} \quad (3)$$

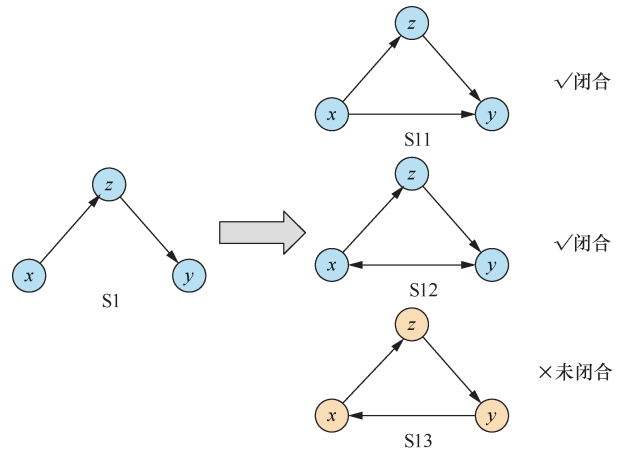


图 4 有向网络中三元组结构闭合示例

基于此, 新的方法 (PTI, potential triad index) 定义如下。

$$s(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} w(x, y, z), \text{pattern}(x, y, z) \in (S1, S2, S3, S4) \quad (4)$$

相似性函数 $s(x,y)$ 是遍历所有的共同邻居节点, 且仅计算其中 S1-S4 结构再求和得到的。可以看出新方法的计算复杂度为 $O(n^2)$, 与 CN 在同一个量级, 其中, n 表示网络中节点的数目。

相似性计算算法流程如下。

```

1) input: directed graph G
2) for each Vertex  $x \in G$  do
3)   for each Vertex  $y \in G$  do
4)     for each Vertex  $z \in G$  do
5)       if  $z \in x.neighbors()$ 
          and  $z \in y.neighbors()$ 
6)         switch(parten(x,y,z)):
7)           case S1:  $s(x,y) += TCI_{S1} * f(z)$ 
8)           case S2:  $s(x,y) += TCI_{S2} * f(z)$ 
9)           case S3:  $s(x,y) += TCI_{S3} * f(z)$ 
10)          case S4:  $s(x,y) += TCI_{S4} * f(z)$ 
11)         end if
12)       end for
13)     end for
14)   end for

```

3 仿真实验

3.1 数据集

本文选择的 9 个公开网络数据集如下。

1) 电子邮件网络 (E-mail) [17]: 欧洲研究机构的电子邮件网络, 节点代表用户, 有向边代表用户发送过邮件。

2) 论文引用网络 (Kohonen) [18]: Kohonen network 也被称为自组织映射 (SOM), 是一种展示和分析多维数据的方法。该网络是与 SOM 相关的论文引用网络。

3) 论文引用网络 (SmaGri) [18]: 与 Small 等相关的论文引用网络。

4) 食物链网络 (FWMW) [19]: 红树林河口湿季的食物链网络, 包含 97 种生物、1 492 条

有向边。

5) 线虫代谢网络 (CElegans) [20]: 该数据集是秀丽隐杆线虫的代谢网络。节点是代谢物 (如蛋白质), 连边是代谢物之间的相互作用。

6) 政治博客网络 (PB) [21]: 这是美国政治博客之间的超链接网络。对于博客 A 和 B, 由 A 指向 B 的有向边表示同方向的超链接。原网络是含有自环的和多连边的, 本文会忽略这些特殊情况。

7) 论文引用网络 (Scimet) [18]: 以“科学计量学”为主的论文引用网络。

8) 维基百科 (Wikivote) [22]: 这是维基百科中用户选举管理员的投票网络。节点代表用户, 边代表投票。原网络的连边是有正负边两种情况的, 本文也对其进行归一化处理。

9) Gnutella 网络 (Gnutella) [23]: Gnutella 是一种文件共享网络, 节点表示网络拓扑中的主机, 连边表示主机的连接关系。

表 1 列出了 9 个网络中 4 种三元组结构的样本数以及闭合指数, 可以看出, 三元组 S4 在所有三元组中闭合指数最低, S1 的闭合指数最高。不同网络的闭合指数相差较大, E-mail 网络中 4 种三元组指数比例差不多, SmaGri 网络中 S1 闭合指数最高, Gnutella 网络 4 种结构闭合指数都很低。由此可以看出, 不同网络的演化规律可能有很大差别。此外, 出现频次和闭合指数并没有直接的关系, 出现频次最高的 S2 和 S3 闭合指数并不是最高的。

表 1 最后一行表示 4 种三元组结构在 9 个网络中的平均闭合指数, 为了进一步简化计算, 不必每次统计整个网络的三元组闭合情况, 同时仍能够保持较好的预测精度, 建议 PCI 写为

$$PCI_{xyz} = \begin{cases} \frac{3}{8}, & xyz \in S1 \\ \frac{1}{4}, & xyz \in S2, S3 \\ \frac{7}{8}, & xyz \in S4 \end{cases} \quad (5)$$

表 1 数据集中 4 种三元组结构及闭合指数

名称	S1 数量	闭合指数	S2 数量	闭合指数	S3 数量	闭合指数	S4 数量	闭合指数
E-mail	1 173 174	0.229 8	1 099 434	0.245 2	1 374 856	0.196 0	1 173 174	0.213 5
Kohonen	96 893	0.071 0	64 724	0.163	1 062 268	0.006 5	96 893	0
SmaGri	33 556	0.129 4	45 576	0.095 2	194 988	0.022 3	33 556	0
FWMW	16 010	0.353 6	35 242	0.160 6	26 372	0.214 7	16 010	0.134 4
CElegans	19 783	0.163 4	38 832	0.083 3	23 802	0.135 8	19 783	0.048 4
PB	503 959	0.247 5	1 244 750	0.099 4	697 328	0.178 9	503 959	0.094 7
Scimet	49 539	0.096 0	94 560	0.050 3	129 984	0.036 6	49 539	0.000 1
Wikivote	3 675 208	0.146 1	6 953 586	0.077 2	11 413 432	0.047 0	3 675 208	0.025 4
Gnutella	114 608	0.007 1	256 050	0.003 2	234 234	0.003 5	114 608	0.000 5
平均值	631 414	0.160 4	1 092 528	0.101 8	1 684 140	0.100 5	631 414	0.057

称将式(5)固定闭合概率代入式(4)得到的相似度为 FPTI (fixed potential triad index)。

3.2 评价指标

实验度量指标采用 AUC (area under the receiver operating characteristic curve) 指标^[24]。其具体含义是正确预测测试集中边的值大于不存在边的概率。AUC 的计算通常采用随机抽样的方式进行估计。每次随机从测试集中选取一条边, 并随机选一条不存在边, 如果测试集中的边分数值较大, 就加 1 分, 相等就加 0.5 分。AUC 可以计算为

$$AUC = \frac{n_1 + 0.5n_2}{n} \quad (6)$$

其中, n_1 表示测试集中的边分数大于不存在的边分数的次数, n_2 代表分数值相等的次数。

3.3 基准方法

LibenNowell 等在链路预测方面所做的开创性工作引起了学术界对于该问题的重视, 其提出的一些方法也成为链路预测研究中进行性能比较的基准算法^[25], 如 SO^[26]、HPI^[27]、HDI^[27]、LHN^[28]等。时至今日, 这些方法仍被大量学者使用, 如文献[11-12]。为了将这些基

准方法应用到有向网络, 本文参照文献[29]中的方法将其进行有向化。表 2 给出了基准方法的定义以及推广到有向网络中的形式。在无向网络中, $\Gamma(x)$ 表示节点 x 的邻居节点集合, $|\Gamma(x)|$ 表示节点 x 的度数; 在有向网络中, 以 x 为源节点的邻居节点集合记为 $\Gamma_{\text{out}}(x)$, x 的出度记为 $|\Gamma_{\text{out}}(x)|$, 以 x 为目的节点的邻居节点集合记为 $\Gamma_{\text{in}}(x)$, x 的入度记为 $|\Gamma_{\text{in}}(x)|$ 。 x 的出入度之和记为 $|\Gamma_{\text{all}}(x)|$ 。

3.4 预测结果

PTI 关注 3.3 节中提到的评价指标 AUC, 并与表 2 中扩展后的 9 种相似性指标进行对比。每个结果都是 100 次独立实验结果的平均值, 每次独立实验都对应一个随机的训练集和测试集。每个数据集中最优的结果加阴影以突出显示。

从表 3 的预测结果可以看出, 对于同一个网络, 前 9 种基准指标的预测精度相差不大, 而对于不同的网络, 预测精度显著不同。一般而言, 如果一个网络具有较高的集聚系数, 那么基于局部相似性的预测指标将得到较好的预测的结果。例如, 效果最差的 Gnutella 网络, 其集聚系数仅有 0.007 2, 这意味着没有多少邻居节点进行相似

表 2 基于局部的链路预测算法

名称	无向网络	有向网络
CN	$s(x, y) = \Gamma(x) \cap \Gamma(y) $	$s(x, y) = \Gamma_{out}(x) \cap \Gamma_{in}(y) $
AA	$s(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log \Gamma(z) }$	$s(x, y) = \sum_{z \in \Gamma_{out}(x) \cap \Gamma_{in}(y)} \frac{1}{\log \Gamma_{out}(z) }$
RA	$s(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{ \Gamma(z) }$	$s(x, y) = \sum_{z \in \Gamma_{out}(x) \cap \Gamma_{in}(y)} \frac{1}{ \Gamma_{out}(z) }$
JA	$s(x, y) = \frac{ \Gamma(x) \cap \Gamma(y) }{ \Gamma(x) \cup \Gamma(y) }$	$s(x, y) = \frac{ \Gamma_{out}(x) \cap \Gamma_{in}(y) }{ \Gamma_{out}(x) \cup \Gamma_{in}(y) }$
SA	$s(x, y) = \frac{ \Gamma(x) \cap \Gamma(y) }{\sqrt{ \Gamma(x) \times \Gamma(y) }}$	$s(x, y) = \frac{ \Gamma_{out}(x) \cap \Gamma_{in}(y) }{\sqrt{ \Gamma_{out}(x) \times \Gamma_{in}(y) }}$
SO	$s(x, y) = \frac{ \Gamma(x) \cap \Gamma(y) }{ \Gamma(x) + \Gamma(y) }$	$s(x, y) = \frac{ \Gamma_{out}(x) \cap \Gamma_{in}(y) }{ \Gamma_{out}(x) + \Gamma_{in}(y) }$
HPI	$s(x, y) = \frac{ \Gamma(x) \cap \Gamma(y) }{\min\{ \Gamma(x) , \Gamma(y) \}}$	$s(x, y) = \frac{ \Gamma_{out}(x) \cap \Gamma_{in}(y) }{\min\{ \Gamma_{out}(x) , \Gamma_{in}(y) \}}$
HDI	$s(x, y) = \frac{ \Gamma(x) \cap \Gamma(y) }{\max\{ \Gamma(x) , \Gamma(y) \}}$	$s(x, y) = \frac{ \Gamma_{out}(x) \cap \Gamma_{in}(y) }{\max\{ \Gamma_{out}(x) , \Gamma_{in}(y) \}}$
LHN	$s(x, y) = \frac{ \Gamma(x) \cap \Gamma(y) }{ \Gamma(x) \times \Gamma(y) }$	$s(x, y) = \frac{ \Gamma_{out}(x) \cap \Gamma_{in}(y) }{ \Gamma_{out}(x) \times \Gamma_{in}(y) }$

表 3 AUC 结果对比

名称	CN	AA	RA	JA	SA	SO	HPI	HDI	LHN	PTI	FPTI
E-mail	0.948 4	0.947 7	0.921 5	0.933 5	0.943 6	0.941 7	0.926 6	0.936 2	0.888 8	0.952	0.952 4
Kohonen	0.925 3	0.894 5	0.925 1	0.925 1	0.925	0.925 3	0.925 4	0.929 4	0.924 9	0.948 7	0.947 1
SmaGri	0.9085	0.883 4	0.907 9	0.905 7	0.906 4	0.905 8	0.907 6	0.905 6	0.905	0.945 5	0.937 7
FWMW	0.771 7	0.743 1	0.677 4	0.728 5	0.696 9	0.694 6	0.699 9	0.694 7	0.6163	0.736 3	0.717 8
CElegans	0.802 9	0.791 9	0.788 3	0.788 7	0.795	0.791 9	0.798 7	0.790 4	0.7861	0.876 9	0.868
PB	0.927	0.92	0.895 4	0.907	0.906 8	0.907 5	0.900 5	0.907 5	0.8783	0.941 2	0.940 2
Scimet	0.894 1	0.873	0.894 1	0.893 9	0.893 9	0.894 1	0.893 8	0.893 9	0.893 9	0.925 3	0.924 1
Wikivote	0.924 1	0.923 9	0.922 3	0.921 3	0.920 9	0.921 6	0.92	0.921 9	0.916 8	0.953 4	0.950 6
Gnutella	0.513 7	0.513 4	0.514	0.513 7	0.513 9	0.513 9	0.513 9	0.513 7	0.513 9	0.529 2	0.529 2

度计算，自然预测精度不高。总体来说，PTI 表现最好。除了 FWMW 食物链网络，PTI 在其余 8 个网络中均取得了最高的预测精度。而在 FWMW 网络中，PTI 的预测精确度也和表现最优的 CN 非常接近。可以看到，除了 FWMW 食物链网络和 Gnutella 点对点网络以外，PTI 在其他

网络的 AUC 都达到了 0.85 以上，在 E-mail 和 Wikivote 网络甚至达到了 0.95 以上。此外，在无向网络表现较好的 RA 和 AA^[2]在有向网络中并没有体现出明显优势。这说明简单地将无向网络预测算法应用到有向网络中，并不完全适应有向网络的一些特性，如有向网络中，每个节点有出

度和入度的不同, 无向网络中的算法会混淆这些特性, 这可能是无向网络中算法在有向网络中区分度不大且表现不是很好的原因之一。

表 3 中 FPTI 代表固定闭合概率的方法, 可以看出, 简化后的方法与原始方法的 AUC 相差很小, 但每次计算时不用统计整个网络的三元组闭合指数, 降低了计算复杂度。在对 AUC 要求不是非常苛刻的场合, 可以考虑使用 FPTI 代替 PTI 使用。

4 结束语

链路预测作为数据挖掘的方向之一, 近几年发展迅猛, 学者对于无向网络进行了系统全面深入的研究, 而有向网络的研究尚处于起步阶段。本文在这方面做出了一些尝试, 针对有向网络特点提出了基于三元组的链路预测方法, 并引入了三元组闭合指数进行相似度计算。在 9 个真实网络的实验中, 新方法在 AUC 上基本都优于基准方法。这说明基于三元组的预测方法更适合有向网络, 能够在一定程度上体现有向网络的连边机理。然而, 本文仅考虑了 9 种三元组中的 4 个, 其余 5 种结构对于相似度的影响未详细分析, 考虑并不全面。同时, 本文仅关注了有向网络, 对于加权网络或者动态网络的分析是下一步的工作。

参考文献:

- [1] LYU L, ZHOU T. Link prediction in complex networks: a survey[J]. *Physica A: Statistical Mechanics And Its Applications*, 2011, 390(6): 1150-1170.
- [2] 吕琳媛, 周涛. 链路预测[M]. 北京: 教育出版社, 2013.
- [3] CLAUSET A, MOORE C, NEWMAN M E J. Hierarchical structure and the prediction of missing links in networks[J]. *Nature*, 2008, 453: 98-101.
- [4] GUIMERA R, SALES-PARDO M. Missing and spurious interactions and the reconstruction of complex networks[J]. *Proc Natl Sci Acad USA*, 2009, 106(52): 22073-22078.
- [5] ADAMIC L A, ADAR E. Friends and neighbors on the Web[J]. *Social Networks*, 2003, 25(3): 211-230.
- [6] SALTON G, MCGILL M J. Introduction to modern information retrieval[M]. Auckland: MuGraw-Hill, 1983.
- [7] JACCARD P. Etude comparative de la distribution florale dans une portion des Alpes et des Jura[J]. *Bulletin de la Société Vaudoise des Science Naturelles*, 1901, 37: 547-579.
- [8] ZHOU T, LYU L, ZHANG Y C. Predicting missing links via local information[J]. *EurPhys J B*, 2009, 71(4): 623-630.
- [9] KATZ L. A new status index derived from sociometric analysis[J]. *Psychometrika*, 1953, 18(1): 39-43.
- [10] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine[J]. *ComputNetw& ISDN Syst*, 1998, 30(1-7): 107-117.
- [11] NARAYANAN A, SHI E, RUBINSTEIN B I P. Link prediction by de-anonymization: how we won the kaggle social network challenge[C]//The 2011 International Joint Conference on Neural Networks (IJCNN). 2011: 1825-1834.
- [12] CORLETTE D, SHIPMAN F M. Link prediction applied to an open large-scale online social network[C]//The 21st ACM Conference on Hypertext and Hypermedia. 2010: 135-140.
- [13] LICHTENWALTER R N, LUSSIER J T, CHAWLA N V. New perspectives and methods in link prediction[C]//The 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2010: 243-252.
- [14] ZHANG Q M, LYU L, WANG W Q, et al. Potential theory for directed networks[J]. *PloS One*, 2013, 8(2): e55437.
- [15] AGHABOZORGI F, KHAYYAMBASHI M R. A new similarity measure for link prediction based on local structures in social networks[J]. *Physica A: Statistical Mechanics and its Applications*, 2018, 501: 12-23.
- [16] BÜTÜN E, KAYA M, ALHAJJ R. Extension of neighbor-based link prediction methods for directed, weighted and temporal social networks[J]. *Information Sciences*, 2018.
- [17] YIN H, BENSON A R, LESKOVEC J, et al. Local higher-order graph clustering[C]//The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 555-564.
- [18] YAO Y, ZHANG R, YANG F, et al. Link prediction in complex networks based on the interactions among paths[J]. *Physica A: Statistical Mechanics and its Applications*, 2018.
- [19] BAIRD D, LUCZKOVICH J, CHRISTIAN R R. Assessment of spatial and temporal variability in ecosystem attributes of the st marks national wildlife refuge, apalachee bay, florida[J]. *Estuarine, Coastal and Shelf Science*, 1998, 47(3): 329-349.
- [20] WATTS D J, STROGATZ S H. Collective dynamics of "small-world" networks[J]. *Nature*, 1998, 393(6684): 440.
- [21] ADAMIC L A, GLANCE N. The political blogosphere and the 2004 US election: divided they blog[C]//The 3rd International Workshop on Link discovery. 2005: 36-43.
- [22] LESKOVEC J, HUTTENLOCHER D, KLEINBERG J. Predicting positive and negative links in online social networks[C]//The 19th

- International Conference on World Wide Web. 2010: 641-650.
- [23] LESKOVEC J, KLEINBERG J, FALOUTSOS C. Graph evolution: densification and shrinking diameters[J]. ACM Transactions on Knowledge Discovery from Data (TKDD). 2007, 1(1): 2.
- [24] HANELY J A, MCNEIL B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve[J]. Radiology, 1982, 143: 29-36.
- [25] 吴祖峰, 梁棋, 刘岍, 等. 基于 AdaBoost 的链路预测优化算法[J]. 通信学报, 2014, 35(3): 116-123.
- WU Z F, LIANG Q, LIU Q, et al. Modified link prediction algorithm based on AdaBoost [J]. Journal on Communications, 2014, 35(3): 116-123.
- [26] SORENSEN T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons[J]. BiolSkr, 1948, 5(4): 1-34.
- [27] RAVASZ E, SOMERA A L, MONGRU D A, et al. Hierarchical organization of modularity in metabolic networks[J]. Science, 2002, 297(5586): 1553-1555.
- [28] LEICHT E A, HOLME P, NEWMAN M E J. Vertex similarity in networks[J]. Phys Rev E, 2006, 73: 026120.
- [29] DEVI S J, SINGH B. Analysis of link prediction in directed and weighted social network structure[C]//The International Symposium on Intelligent Systems Technologies and Applications. 2017: 1-13.

[作者简介]



常圣（1988- ），男，河南郑州人，国家数字交换系统工程技术研究中心硕士生，主要研究方向为链路预测。



马宏（1968- ），男，江苏东台人，国家数字交换系统工程技术研究中心研究员，主要研究方向为社会网络分析、电信网关防护。



刘树新（1987- ），男，山东潍坊人，博士，国家数字交换系统工程技术研究中心助理研究员，主要研究方向为复杂网络、网络信息挖掘。