基于投影的二分网络链接预测

高 曼[1] 陈 崚[1,2] 徐永成[1]
(扬州大学信息工程学院 扬州 225009)[1]
(南京大学计算机软件新技术国家重点实验室 南京 210093)[2]

摘 要 提出基于投影的二部网络链接预测算法。算法首先将二部图投影为一个单部图,在此基础上定义了潜在边的概念,使得对二分网络链接的预测仅在潜在边中进行,大大降低了预测算法的复杂度。定义了潜在边所覆盖的模式以及模式的权重,通过潜在边所覆盖的模式的权重来计算潜在边的可信度,并将其作为该潜在边上存在实际链接的评分。实验结果表明,所提算法能够有效地提高链接预测的速度和结果的精度。

关键词 二部网络,链接预测,投影,潜在边
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.2.027

Projection Based Algorithm for Link Prediction in Bipartite Network

GAO Man[1] CHEN Ling[1,2] XU Yong-cheng[1]
(College of Information Engineering, Yangzhou University, Yangzhou 225009, China)[1]
(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)[2]

Abstract An algorithm for link prediction in a bipartite network was presented. In the algorithm we first mapped the bipartite network to unipartite one called projected graph. Based on the projected graph, we defined the concept of potential link. We performed the link prediction only within the potential links so as to reduce the computation time. We also defined the pattern covered by the potential links and the weight of the patterns. By calculating the weight of the patterns a potential link covers, the confidence of the potential link can be obtained, which can be used as the final score of link prediction. Experimental results show that our algorithm can get faster speed and higher quality of link prediction results.

Keywords Bipartite network, Link prediction, Projection, Internal links

1 引言

现实世界中存在的关系已越来越多地被抽象成各种复杂的网络,例如在线社交网络、蛋白质相互作用网络、神经网络、电力网络、航空网络、用户商品网络等等。这些网络中包含了成千上万的节点以及节点之间的连边。复杂网络作为复杂系统的一种拓扑近似,在构建过程中,由于时间和空间或者实验条件的限制,难免有错误或冗余的链接出现,还有不少潜在的链接并未探测到。再者,复杂网络往往是随时间动态演化的,其链接会不断地添加或去除。因此,需要根据已知的网络信息对缺失的链接以及未来的链接进行预测,这就是网络链接预测问题[1-3]。

链路预测问题有重大的实际应用价值。如在生物领域研究中,蛋白质相互作用网络和新陈代谢网络[4]节点之间存在链接,即存在相互作用关系。但揭示该类网络中隐而未现的相互作用关系需要耗费高额的实验成本,而链路预测方法的结果可以指导实验,提高实验的成功率,从而降低实验成本。对疾病-基因网络的丢失和可疑链接预测研究,有助于探索疾病的发生机制,预测和评价相应的治疗手段,同时还可以寻找新的药物靶标,为新药研发开辟新的途径[5]。

在社会网络分析中,链路预测同样可以作为准确分析社会网络结构的有力的辅助工具。如近几年在线社交网络发展非常迅速,链路预测可以将用户潜在的朋友推荐给用户[6]。在社会关系分析中,可以发现人际之间潜在的联系[7,8]。链路预测的思想和方法还可以用于学术网络中判断一篇学术论文的类型以及合作者[9]。链路预测的方法也可直接用于信息的推荐,在电子商务中可以用于对客户的商品推荐[10],还可以用于电子邮件的预测[11],在无线通讯网络中判断一个手机用户是否产生了切换运营商的倾向。在对犯罪分子组成的网络的监控中,需要利用链接预测来发现犯罪分子间隐藏的联系,以防止犯罪或恐怖活动的发生。

链接预测研究不仅具有广泛的实际应用价值,也具有重要的理论研究意义。例如,链路预测的研究也可以从理论上帮助人们认识复杂网络演化的机制[12]。由于刻画网络结构

到稿日期:2015-01-13 返修日期:2015-06-07 本文受国家自然科学基金(61379066,61070047,61379064,61472344),国家973项目(2012CB316003),江苏省自然科学基金(BK20130452,BK2012672,BK2012128,BK20140492),江苏省教育部门自然科学基金(12KJB520019,13KJB520026),江苏省研究生培养创新工程项目(CXZZ13_0173)资助。

高 曼(1991-),女,硕士生,主要研究方向为数据挖掘、人工智能,E-mail:15396768192@163.com;陈 崚(1967-),男,教授,主要研究方向为数据挖掘、并行与分布式处理、人工智能;徐永成(1989-),男,硕士生,主要研究方向为数据挖掘、人工智能。

特征的统计量非常多，很难比较不同的机制孰优孰劣。链路预测可以为演化网络机制提供一个简单统一且较为公平的比较平台，从而推动复杂网络演化模型的理论研究。

对网络链接的预测有基于节点相似性、基于最大似然估计、基于概率模型等方法。基于节点相似性方法假设两个节点之间相似性越大，它们之间存在链接的可能性就越大。为此，有很多有关节点的相似性的定义，如基于局部信息的相似性指标，其主要有共同邻居(CN)指标、Salton 指标[13]、Jaccard 指标[14]、Sorenson 指标[15]、大度节点不利指标(HDI)[16]、大度节点有利指标(HPI)和 LHN-I 指标[17]、优先链接指标(PA)[18]、Adamic/Adar 指标[19]和 RA 指标[20]等。基于路径的相似性指标主要有局部路径指标(LP)[21]、LHN-II[17]指标和 Katz 指标[22]等。基于随机游走的相似性指标包含平均通勤时间(average commute time)[23]、Cos＋指标[24]、重启的随机游走(random walk with restart)[25]、SimRank 指标[26]等。周涛、吕琳媛和张翼成[27,28]提出了两种新指标：资源分配指标和局部路径指标。这两种指标具有明显好于包括 Adamic/Adar 指标在内的 9 种已知指标的预测能力。刘伟平和吕琳媛[29]提出了两种基于网络局部随机游走的相似性指标，通过与其他 5 种相似性指标的比较，发现有限步的随机游走可以给出比全局收敛后的预测精度更好的结果。饶君等人[30]针对 9 种局部信息的相似性指标，设计并实现了一种基于 MapReduce 计算模型的并行链接预测算法，使链接预测应用于大型复杂网络。东昱晓等人[31]提出了一种基于节点引力指数的预测算法，在保持低时间复杂度的同时，提高了预测的准确率。

链路预测的另一类方法是基于最大似然估计的。该方法将网络的链接看作某种内在的层次结构，或者是随机分块模型结构，相应提出了层次结构模型和随机分块模型，利用最大似然估计的算法进行链接预测。层次结构模型链接预测方法在处理具有明显层次组织的网络，如恐怖袭击网络和草原食物链时，具有较好的精确度。但是，由于每次预测要生成很多个样本网络，因此其计算复杂度非常高，只能处理规模不太大的网络。基于随机分块模型的链路预测方法不仅可以预测缺失边，还可以预测网络的错误链接，例如纠正蛋白质相互作用网络中的错误链接。

还有一类链接预测的方法是基于概率模型的，该类方法的基本思路就是建立一个含有一组可调参数的模型，然后使用优化策略寻找最优的参数值，使得所得到的模型能够更好地再现真实网络的结构和关系特征，网络中两个未相连的节点间存在链接的概率就等于在该组最优参数下它们之间产生链接的条件概率。概率模型的优势在于有较高的预测精确度，但是计算的复杂度以及非普适性的参数使其应用范围受到限制。

二分网络是复杂网络中的一种重要的表现形式，由两部分不同类型的顶点构成，同一类型部分的两个顶点不相连。现实世界中的许多网络，都呈现出自然的二分结构，比如作者与他们所发表的论文著作形成的作者-论文合作网[36,37]、演员与他们所演出的电影作品形成的演员-事件合作网[38]、投资者与他们持有股份的公司之间形成的股份网络[39,40]、疾病-基因网络[41]、俱乐部成员与活动网络[42]、观众与歌曲网络[43]、在 P2P 系统中计算机终端数据网络[44]等。从这些二分网络中挖掘潜在的链接，可以帮助我们更好地认识所处的世界。例如，通过对在线社交网络的研究，可以预测出哪些尚未成为朋友的用户在将来可能成为朋友，这可以帮助我们在自己的好友圈里找到“失散”或“失联”的好友，也可以让我们认识我们朋友的朋友，从而扩大我们的交友圈。对二分网络研究的结果可以用到用户-商品推荐等更多的系统中，基于此，越来越多的研究者开始将目光转向二分网络的研究。

本文提出一种基于投影的二部网络链接预测算法。算法首先将二部图映射到一个投影图上，在此基础上定义了潜在边的概念，并通过该投影图来检测原二部图中的潜在边。进而通过对潜在边计算权重来预测他们在二部网络中出现的可能性。实验结果表明，该算法能够有效地提高链接预测的速度和结果的精度。

2 二部网络及其投影图

二部网络可以用一个无向简单二部图 $G=(U,V,E)$ 来表示，这里，U 和 V 分别是 G 的两部分顶点的集合，E 为 G 的边的集合。对于任意边 $(u,v)\in E$，必有 $u\in U$、$v\in V$，即只有不同部分间的顶点才能连接，而在同一部分的顶点之间不存在链接。例如，图 1 所示为一个二分网络，在该网络中，三角形顶点为同一类型的顶点，圆形顶点为另一类同类型顶点，同类型顶点之间无边相连。

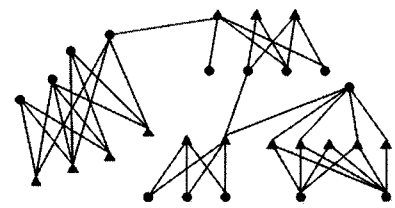


图 1 二分网络示意图

二部分网络中 U 部分节点 u 的邻居集合定义为 $N(u)=\{v|v\in V,(u,v)\in E\}$。即节点 u 的邻居是在二部分网络中所有与 u 相连的顶点的集合。类似地，我们可以对 V 部分节点 v 定义其邻居集合 $N(v)$。

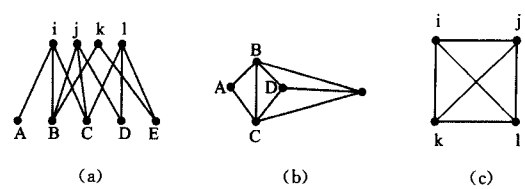
为了分析二部网络中的潜在链接，我们将二部网络的某一部分的顶点投影到一个单部图中。

定义 1(投影图) 简单二部图 $G=(U,V,E)$ 的在顶点集合 U 上的投影图为一个单部图 $G_u=(U,E_u)$，G_u 顶点的集合为原二部图 G 中 U 部分的顶点，G_u 的边的集合为：

$$E_u=\{(u,w)|u,w\in U,\exists v\in V,v\in N(u)\cap N(w)\} \quad (1)$$

类似地，也可以定义二部图 $G=(U,V,E)$ 的在顶点集合 V 上的投影图 G_v。由定义 1 可见，U 部分的顶点 u、w 如果在二部图 $G=(U,V,E)$ 至少有一个共同邻居，则在投影图中就有相应的边 (u,w)。

例如，将如图 2(a) 所示的二部图按照底部节点进行投影，可得到如图 2(b)所示的投影图。将其按照顶部节点进行投影，可得到如图 2(c)所示的投影图。



(a) (b) (c)

图 2 二部图 G 及其投影图

3 潜在边及其可信度

为了对二部网络进行链接预测，我们通过它的投影图来定义潜在边的概念。

定义 2(潜在边) 设二部图 $G=(U,V,E)$ 中，有顶点对 (u,v)，这里 $u\in U$、$v\in V$ 且 $(u,v)\notin E$。设 $G_u=(U,E_u)$ 为对二部图 G 的 V 部分的顶点投影后得到的投影图。设在二部图 G 中加入边 (u,v) 而得到二部图 G'，并设将 G' 的 U 部分的顶点投影后得到投影图 $G'_u=(U,E_u')$。若 $G_u=G_u'$，则在二部图 G 中的这条未连接的边 (u,v) 就是一条潜在边。

换句话说，潜在边就是这样的一对节点 (u,v) 连接起来的边：它的加入不会对原二部图的投影产生变化。

根据以上定义，可以发现图 2(a)中的未连接的顶点对 (B,l) 就是一条潜在边。事实上，二部图 G 中节点 l 的邻居为 $N(l)=\{C,D,E\}$，而在图 2(b)和图 2(c)所示的投影图中，这些点已经与 B 连接了，所以加入边 (B,l) 到二部图 G 中不会对投影图产生影响。

设二部图 $G=(U,V,E)$ 中，设有顶点对 (u,v)，$u\in U$、$v\in V$ 且 $(u,v)\notin E$，设对二部图 G 的 U 部分的顶点投影后得到投影图 $G_u=(U,E_u)$。记 u 在投影图 $G_u=(U,E_u)$ 中的邻居节点集合为 $N_u(u)=\{w|w\in U,(u,w)\in E_u\}$，记 v 在对二部图 $G=(U,V,E)$ 中的邻居节点集合为 $N(v)=\{w|w\in U,(u,v)\in E\}$，根据潜在边的定义，易知 (u,v) 为潜在边的条件是：

$$N_u(u)\cap N(v)\neq\Phi \text{ 且 } u\notin N(v) \tag{2}$$

集合 $N_u(u)\cap N(v)$ 为同时在二部图中与 V 中的顶点 v 相邻接、同时又在投影图中和顶点 u 相邻接的顶点集合，其中的每一个顶点与 u 在投影图中构成一条边。设 (u,v) 为潜在边，由于 $u\notin N(v)$，$(u,v)\notin E$，即二部图中不存在边 (u,v)。但如果在二部图中加入边 (u,v)，在投影图上应增加的任何边 (u,w)，都会有 $w\in N_u(u)\cap N(v)$。因为 $w\in N_u(u)$，则有 $(u,w)\in E_u$，即应该在投影图增加的边 (u,w) 原来就存在，因而新的投影不会改变。我们称投影图中这样的边 (u,w) 为潜在边 (u,v) 所覆盖的一个模式。

定义 3(潜在边覆盖的模式) 设 (u,v) 为潜在边，对任意顶点 $w\in N_u(u)\cap N(v)$，投影图中的边 (u,w) 为潜在边 (u,v) 所覆盖的一个模式。

潜在边在投影图上覆盖了一个或多个模式，它们如果变为真实的边，由于有已经存在的类似模式，说明它们潜在的可能性较大。因为增加潜在边为真实的边只会产生与原有图中相似的链接模式，通过图中的潜在边来进行二部图上的链接预测。例如，假设图 2(a)中的底部顶点表示物品，上部顶点表示顾客。边 (A,i) 的存在表示顾客 i 购买了物品 A。如果顾客 i 同时购买物品 A 和 B，我们称 $\{A,B\}$ 构成了一种购物模式。在图 2(a)中，如果顾客 i 购买了物品 B，其产生的所有模式都有已存在的模式与之相同，因此顾客 i 购买物品 B 的可能性较大，即在图 2(a)中的二部图上出现边 (B,l) 的可能性较大。反过来，如果一个未连接边不是潜在边，则找不到所覆盖的模式，说明它们潜在的可能性极小。因此可以仅在潜在边之中寻找可能的链接。从而，可以大大减少链接预测中的候选边范围，提高链接预测算法的复杂度。

我们采用相似度的方法进行链接预测，即对每个潜在边进行评分，以表示该潜在边出现的可能性。因此，我们用潜在边所覆盖的模式个数来衡量其出现的可能性，我们称之为潜在边的可信度。为了反映潜在边出现的可能性，我们在对于可信度的计算中考虑了潜在边所覆盖的模式数和这些模式的权重。

对于一个潜在边来说，集合 $N_u(u)\cap N(v)$ 的大小即为边 (u,v) 加入二部图 G 后，已有的相同的模式的个数。$|N_u(u)\cap N(v)|$ 越大，说明顶点对 (u,v) 存在链接的可能性越大。

例如在图 3 所示的二部图中，虽然 (A,i) 和 (A,j) 都是潜在边，但它们覆盖的模式的个数不一样。(A,i) 仅覆盖了模式 $\{A,B\}$，而 (A,j) 覆盖了模式 $\{A,B\}$、$\{A,C\}$、$\{A,D\}$。显然，(A,j) 出现链接的可能性要比 (A,i) 大。因此，我们用潜在边所覆盖的模式个数来衡量其出现的可能性，称之为潜在边的可信度。

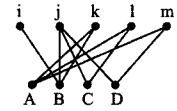


图 3 二部图 G 的潜在边 (A,i) 和 (A,j)

一条隐含边所覆盖的模式，对应了投影图上的一条边，在投影图上连接的两点只说明了这两个节点在原图中有共同邻居，而不能说明它们共同邻居的多少，为了在投影过程中保留原图更多的信息，我们采用了权值投影，即给投影图 G_u 中的每条边 (A,B) 赋予权值，来反映顶点 A、B 在原图 G 上的相似程度。

首先应该考虑 A、B 在原图 G 上的共同邻居数。例如，图 4(a)及图 4(b)中的二部图的投影图 G_u 皆为图 4(c)所示。底部顶点 A、B 投影后，在投影图 G_u 中皆为边 (A,B)。但图 4(a)中的顶点 A、B 只有 1 个共同邻居，而图 4(b)中 A、B 有 4 个共同邻居，在相应的投影图 G_u 中，边 (A,B) 的权重应该较大。

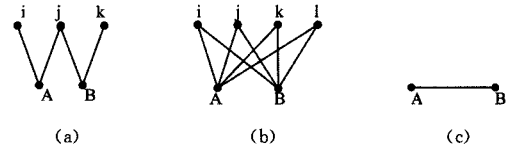
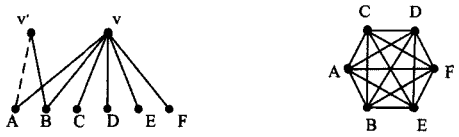


图 4 两个不同的二部图及其在底部的投影

其次，还应该考虑 A、B 共同邻居的权重。如果 A、B 的共同邻居具有较小的度数，则 A、B 的权重较小。如图 5(a)所示的二部图，图 5(b)为其投影图，图 6(a)所示的二部图，图 6(b)为其投影图。在这两个二部图中，(A,v') 皆为潜在边，$\{A,B\}$ 是它们所覆盖的模式，但在图 5(a)所示的二部图中，A、B 的共同邻居 v 的度数为 2，而图 6(a)所示的二部图中，A、B 的共同邻居 v 的度数为 4。我们设 A、B 为两个顾客，顶点 v 为商品，则在图 5(a)中商品 v 仅被 A、B 所共同购买。而图 6(b)中的商品 v 被 A、B 等所有 6 个顾客共同购买，显然图 5(a)中的顾客 A、B 的购物相似度较高，因而图 5(b)中的边 (A,B) 的权重应该较大。

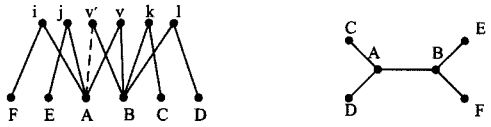
图 5 潜在边 (A,v') 及其投影



(a) (b)

图 6 潜在边(A,v')及其投影

此外,A,B 本身在二部图中的度数也同样影响投影图中边(A,B)的权重。例如在图 5(a)和图 7(a)中的二部图中,(A,v')皆为潜在边。在图 5(a)的二部图中顶点 A、B 的度数分别为 1、2,而在图 7(a)的二部图中,顶点 A、B 的度数分别为 3 和 4。设 A、B 为两个顾客,顶点 v 为商品,在图 5(a)的二部图中,顾客 A 仅购买商品 v,顾客 B 仅购买商品 v 和 v'。而在图 7(a)的二部图中,顾客 A、B 还购买了许多其他的商品,而商品 v 仅是其中唯一的共同购买的商品。显然,图 7(a)中的顾客 A、B 的购物相似度较低,因而投影图 7(b)中边(A,B)的权重应该较小。



(a) (b)

图 7 潜在边(A,v')及其投影

基于以上的分析,我们定义投影图的边上的权重如下:

定义 4(投影图的边上的权重) 设 $G_u=(U,E_u)$ 为二部图 $G=(U,V,E)$ 的投影图。边 $(A,B)\in E_u$ 为 G_u 的一条边,(A,B)上的权重为:

$$w(A,B)=\frac{2}{D(A)+D(B)}\sum_{v\in N(A)\cap N(B)}\frac{1}{D(v)} \tag{3}$$

其中,$D(A)$、$D(B)$、$D(v)$分别为顶点 A、B、v 在二部图中的度,$N(A)$、$N(B)$分别为顶点 A、B 在二部图中邻接顶点的集合。

由式(3)可以看出,如果顶点 A、B 在二部图中的共同邻居越多,且这些共同邻居的度越低,投影图中的边(A,B)的权重就越高;同时,顶点 A、B 在二部图中的度越低,投影图中的边(A,B)的权重也就越高。由于投影图的每一条边代表一个模式,边(A,B)的权重也就是模式$\{A,B\}$的权重。

在定义了权重后,对潜在边的可信度做如下定义。

定义 5(潜在边的可信度) 设(A,a)为二部图 $G=(U,V,E)$的潜在边,(A,a)的可信度定义为:

$$S(A,a)\sum_{\{A,B\}\in\Gamma(A,a)}w(A,B) \tag{4}$$

这里 $\Gamma(A,a)$为潜在边(A,a)所覆盖的模式的集合,定义为:

$$\Gamma(A,a)=\{\{A,B\}\mid B\in N_u(A)\cap N(a)\} \tag{5}$$

由定义 5 可以看出,潜在边(A,a)的可信度为它所覆盖的模式在投影图中相应的边的权重之和。由式(5)可知,如果潜在边(A,a)所覆盖的模式越多,且这些在投影图中相应的边的权重越高,潜在边(A,a)的可信度就越高。由于投影图中相应的边的权重反映了这些模式的相似度,这样定义的潜在边的可信度反映了该潜在边出现的可能性。因此我们用潜在边(A,a)的可信度作为该边链接预测的指标。

4 算法的框架与复杂度分析

根据以上分析,本文算法首先将二部图转换成一个投影图,并按式(3)计算投影图上每一条边的权重,然后对每个潜在边按式(4)计算可信度,作为该边链接预测的指标。算法的框架描述如下。

算法 1 PLP (Potential-Link-Prediction)

```
输入:G=(U,V,E):二部图;
输出:潜在边的集合及其可信度矩阵 S;
Begin
  (1)/*构造投影图 Gu'=(U,Eu')*/
     Eu=Φ;
     For every node A in U do
       For every node v in N(A) do
         For every node B in N(v) do
           Eu=Eu∪{(A,B)};
         Endfor
       Endfor
     Endfor
  (2)/*计算投影图上每一条边的权重*/
     For every edge (A,B) in Eu do
       Calculate the weight of edge (A,B) according to (3);
  (3)/*计算每个潜在边的可信度*/;
     For 投影图 Gu 上的每一个顶点 A do
       For A 在 Gu 上的每一个邻接点 B do
         For 在二部图 G 中与 B 相邻的每一个顶点 a do
           If (A,a)∉E then
             S(A,a)=S(A,a)+w(A,B);
           Endif
         Endfor
       Endfor
     Endfor
  (4) 输出 S 中的潜在边的可信度,即为这些边出现的概率。
End
```

设$|U|=m$,d 为 U、V 部中顶点的度的最大值。该算法的步骤(1)三重循环的最大执行次数依次为 m、d、d,因而复杂度为 $O(m\times d^2)$。该算法的步骤(2)计算投影图上每一条边的权重。设 A 为投影图上的一个顶点,由于投影图上与 A 链接的每一条边对应二部图中一条与 A 链接的长度为 2 的路径,而这样的路径的条数不会超过 A 的度数,因此投影图上与 A 链接的边的条数最多为 d。由于投影图上最多有 $|U|=m$ 个顶点,因此投影图上最多有 $m\times d$ 条边。由式(3)易知计算一条边的权重与端点的度相关,因而计算量最多为 d。由此可见,算法步骤(2)的复杂度为 $O(m\times d^2)$。该算法的步骤(3)三重循环的最大执行次数依次为 m、d、d,因而复杂度也为 $O(m\times d^2)$。综上所述,算法 PLP 的时间复杂度为 $O(m\times d^2)$。通常,网络中顶点的最大度 d 可以看成是一个常量,因此算法的复杂度为 $O(m)$,与 U 部顶点的个数呈线性增长。算法 PLP 时间复杂度较低的原因是它仅在潜在边之中寻找可能的链接,缩小了链接预测中的候选边范围,大大降低了计算量。

5 实验结果与分析

为了验证算法 PLP 的效率以及准确率,用 Southern Women、Divorce、Scotland 3 个数据集进行实验,并对结果进行分析。在实验中,用 Matlab 完成整个实验代码的编写过程,用 Pajek、Visio 等工具完成本文中图的绘制。

5.1 对 Southern Women 数据集的实验

为了验证算法 PLP 的预测准确度，实验使用由 Davis 在 1930 年间收集的 Southern Women 数据集，该数据集描述了密西西比州南方女子俱乐部中成员参加活动的情况。因为这个数据集具有明显的社区结构，所以被广泛地用来测试分析。Southern Women 网络可用一个二部图来表示。该二部图的一部分顶点表示妇女，另一部分顶点表示活动。如果一个妇女参加了一个活动，那么在该妇女和这个活动的顶点之间则有一条连边。在该网络中，妇女顶点之间没有链接，活动顶点之间也没有链接。图 8 所示为 Southern Women 网络的结构。在图 8 中，圆形节点表示 18 个妇女，三角形节点表示 14 个活动，连边表示一个妇女加入了一个活动，网络中共有 93 条边。Southern Women 网络的主要结构参数如表 1 所列。

图 8 Southern Women 二分网络

表 1 Southern Women 数据集

节点数⊥	节点个数 T	边个数 E	测试集边数	训练集边数	潜在边数
18	14	93	10	83	68

在该数据集上对算法 PLP 进行了 10 次实验，每次随机抽取 10 条边作为测试集，剩下的边作为训练集。同时对该数据集使用 CN、Katzl 算法进行了测试。我们使用 AUC 作为评价标准比较了算法的精度，结果如表 2 所列。

表 2 Southern Women 数据集上的预测结果的精度比较(AUC)

Southern Women	PLP	CN	Katz
1	0.9293	0.8743	0.9107
2	0.9268	0.8927	0.8997
3	0.9172	0.8859	0.9043
4	0.9452	0.9031	0.9155
5	0.9465	0.9325	0.9239
6	0.9217	0.8791	0.9392
7	0.9484	0.9033	0.9179
8	0.9389	0.8976	0.9007
9	0.9637	0.9077	0.9351
10	0.9274	0.8939	0.9114

由表 2 可以看出，算法 PLP 在 10 次测试中的 AUC 值均大于 CN，有 9 次大于 Katz 算法，这说明算法 PLP 能够取得比其它算法精度更高的预测结果。

5.2 对 Divorce 数据集的实验

Divorce 数据集收集了导致美国 50 个州的居民离婚的 9 大原因。Divorce 网络的结构如图 9 所示。图中的 1 号节点至 50 号节点表示 50 个州，51 号节点至 59 号节点表示离婚的原因。Divorce 网络中有 50 个州参与这项调查，主要的原因归结为 9 类，表 3 给出了该网络的主要参数。

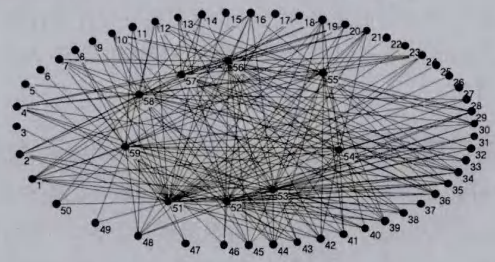


图 9 Divorce 二分网络图

表 3 Divorce 数据集

节点数⊥	节点个数 T	边个数 E	测试集边数	训练集边数	潜在边数
50	9	225	23	202	130

在该数据集上对算法 PLP 进行了 10 次实验，每次随机抽取 10 条边作为测试集，剩下的边作为训练集。同时对该数据集使用 CN、Katzl 算法进行了测试。我们使用 AUC 作为评价标准比较了算法的精度，结果如表 4 所列。

表 4 Divorce 数据集上的预测结果的精度比较(AUC)

Divorce	PLP	CN	Katz
1	0.9413	0.8920	0.9239
2	0.9493	0.9021	0.9085
3	0.9321	0.8874	0.8974
4	0.9452	0.9210	0.9230
5	0.9571	0.9170	0.9184
6	0.9248	0.8970	0.8678
7	0.9770	0.9207	0.9387
8	0.9447	0.8799	0.9149
9	0.9609	0.9309	0.9407
10	0.9545	0.9024	0.9257

由表 4 可以看出，算法 PLP 在 10 次测试中的 AUC 值均大于 CN 和 Katz 算法，这说明算法 PLP 能够取得比其它算法精度更高的预测结果。

5.3 对 Scotland 数据集的实验

我们还采用了 20 世纪初苏格兰连锁企业的数据集进行了测试。该集合收集了苏格兰早期的 108 个公司与 136 位股东之间的关系，每一位股东可能在不同的公司任职，每一家公司也可能有不同的股东。这样公司于股东之间就形成了二分网络的关系。但与 Southern Woman、Divorce 数据集不同，该数据集是非连通图，有许多离群点，如图 10 所示。因此我们首先抽取了该网络的最大连通分量，如图 11 所示。实验在该数据集的最大连通分量上进行。得到的最大通图的结构参数如表 5 所列。

图 10 Scotland 网络图

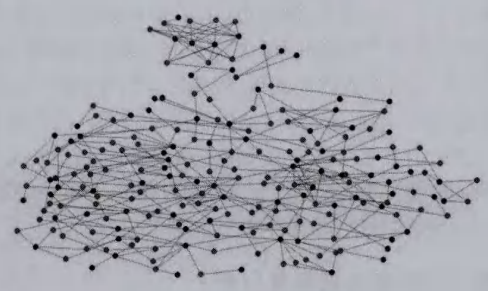


图 11 Scotland 最大连通图

表 5　Scotland 最大连通图的结构参数

节点数 ⊥	节点个数 T	边个数 E	测试集边数	训练集边数	潜在边数
131	86	348	35	313	1423

我们在该数据集上对算法 PLP 进行了 10 次实验，每次随机抽取 10 条边作为测试集，剩下的边作为训练集。同时对该数据集使用 CN、Katzl 算法进行了测试。我们使用 AUC 作为评价标准比较了算法的精度，结果如表 6 所列。

表 6　Scotlan 最大连通图上的预测结果的精度比较(AUC)

Scotland	PLP	CN	Katz
1	0.9179	0.8901	0.9011
2	0.8871	0.8953	0.9108
3	0.9235	0.8989	0.9019
4	0.8813	0.8947	0.8997
5	0.8951	0.8759	0.8796
6	0.8880	0.8645	0.8849
7	0.9136	0.8830	0.8904
8	0.8818	0.8794	0.8997
9	0.9144	0.8793	0.9023
10	0.9545	0.9307	0.9124

由表 6 可以看出，算法 PLP 在 10 次测试中的 AUC 值有 8 次大于 CN，有 7 次大于 Katz 算法，因此 PLP 在较稀疏的网络中仍然可以保持高预测精度。

从表 6 可以明显发现，AUC 值在其它数据集上有普遍的下降，再观察图 11，可以发现 Scotland 图比较稀疏，因此未连接的边在实验中占的比重较前两个数据集大，这是可以理解的。但根据 3 种数据集的实验结果可以发现，我们的算法在二分网络上的预测准确度比其它方法要高。

结束语　本文提出了一个基于投影的二部网络连接预测算法。算法首先将二部网络转换成一个投影图，在此基础上定义了潜在边的概念，以及潜在边所覆盖的模式。通过潜在边所覆盖的模式的权重来计算潜在边的可信度，作为该潜在边上存在实际链接的评分。由于算法使得对二分网络链接预测仅在潜在边中进行，大大减低了预测算法的复杂度。实验结果表明，所提算法能够有效地提高链接预测的速度和结果的精度。

参 考 文 献

[1] Ryan N. Lichtenwalter, New precepts and method in link prediction[C]//Proceedings of ACM KDD'10. 2010:243-252

[2] Lv Lin-yuan, Zhou Tao. Link predication in complex networks: A survey[J]. Physica A, 2011, 390:1150-1170

[3] Lv Lin-yuan. Link prediction on complex networks[J]. Journal of University of Electronic Science and Technology of China, 2010, 39(5):651-661(in Chinese)
吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39(5):651-661

[4] Zhang L, Hu K, Tang Y. Predicting Disease-related Genes by Topological Similarity in Human Protein-protein Interaction Network[J]. Central European Journal of Physics, 2010, 8(4): 672-682

[5] Guimera R, Sales-Pardo M. Missing and Spurious Interactions and the Reconstruction of Complex Networks[J]. Proceedings of the National Academy of Science, USA, 2010, 106(52): 22073-22078

[6] Papadimitriou A, Symeonidis P, Manolopoulos Y. Fast and accurate link prediction in social networking systems[J]. Journal of Systems and Software, 2012, 85(9): 2119-2132

[7] Hossmann T, Nomikos G, Spyropoulos T, et al. Collection and analysis of multi-dimensional network data for opportunistic networking research[J]. Computer Communications, 2012, 35(13): 1613-1625

[8] Jahanbakhsh K, King V, Shoja G C. Predicting missing contacts in mobile social networks[J]. Pervasive and Mobile Computing, 2012, 8(5): 698-716

[9] Sun Y, Barbery R, Gupta M, et al. Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks[C]// Proceedings of 2011 International Conference on Advances in Social Networks Analysis and Mining(ASONAM 2011). 2011: 121-128

[10] Li Xin, Chen Hsin-chun. Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach[J]. Decision Support Systems, 2013, 54(2): 880-890

[11] Huang Z, Lin D K J. The time-series link prediction problem with applications in communication surveillance[J]. INFORMS J. on Computing. 2009, 21: 286-303

[12] Liu H K, Lv L Y, Zhou T. Uncovering the network evolution mechanism by link prediction[J]. Sci. Sin. Phys Mech & Astron, 2011, 41(7): 816-823(in Chinese)
刘宏鲲, 吕琳媛, 周涛. 利用链路预测推断网络演化机制[J]. 中国科学: 物理学力学天文学, 2011, 41(7): 816-823

[13] Salton G, Mcgill M J. Introduction to modern information retrieval[M]. Auckland: MuGraw-Hill, 1983

[14] 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

[15] Sorensen T. A method of establishing groups of equalamplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons [J]. Biol Skr, 1948, 5(4): 1-34

[16] Ravasz E, Somera A L, Mongru D A, et al. Hierarchical organization of modularity in metabolic networks[J]. Science, 2002, 297(5586): 1553-1555

[17] Leicht E A, Holme P, Newman M E J. Vertex similarity in networks[J]. Phys Rev E 73, 2006, 73: 026120

[18] Barabasi A-L, Albert R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512

[19] Adamic L A, Adar E. Friends and neighbors on the web[J]. Social Networks, 2003, 25(3): 211-230

[20] Zhou T, Lv L, Zhang Y C. Predicting missing links via local information[J]. Eur Phys J B, 2009, 71(4): 623-630

[21] Lv L, Jin C H, Zhou T. Similarity index based on local paths for link prediction of complex networks[J]. Phys Rev E, 2009, 80: 046122

[22] Katz L. A new status index derived from sociometric analysis [J]. Psychometrika, 1953, 18(1): 39-43

[23] Klein D J, Randic M. Resistance distance [J]. J Math Chem, 1993, 12(1): 81-95

[24] Fouss F, Pirotte A, Renders J M, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation [J]. IEEE Trans Knowl Data Eng, 2007, 19(3): 355-369

[25] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[J]. Comput Netw & ISDN Syst, 1998, 30(1-7): 107-117

(下转第 154 页)

Partial Fingerprint Identification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(1): 72-87

[2] Jing Xiao-yuan, Li Sheng, Zhang D, et al. Optimal subset-division based discrimination and its kernelization for face and palmprint recognition[J]. Pattern Recognition, 2012, 45(10): 3590-3602

[3] Su Yu, Shan Shi-guang, Chen Xi-lin, et al. [J]. Journal of Software, 2010, 21(8): 1849-1862(in Chinese)
苏煜,山世光,陈熙霖,等. 基于全局和局部特征集成的人脸识别[J]. 软件学报, 2010, 21(8): 1849-1862

[4] Nigam A, Gupta P. Iris recognition using consistent corner optical flow [C]//Proc of 11th Asian Conference on Computer Vision. 2012: 358-369

[5] Lee P J-W, Choi S-S, Moon P B-R. An evolutionary keystroke authentication based on ellipsoidal hypothesis space [C]// Proc of the 9th Annual Conference on Genetic and Evolutionary Computation. 2007: 2090-2097

[6] Ahmed, TraoreIssa A A. Biometric recognition based on free-text keystroke dynamics [J]. IEEE Transactions on Cybernetics, 2014, 44(4): 458-472

[7] Ahmed A A E, Traore I. Detecting computer intrusions using behavioral biometrics[C]// Proc of 3rd Annual Conference on Privacy. Security, 2005: 91-98

[8] Jorgensen Z, Yu T. On mouse dynamics as a behavioral biometric for authentication[C]//Proc of the 6th ACM Symposium on Information, Computer and Communications Security. 2011: 476-482

[9] Gamboa H, Fred A L N, Jain A K. Webbiometrics: User verification via web interaction[C]// Proc of 2007 Biometrics Symposium. 2007: 1-6

[10] Pusara M, Brodley C E. User re-authentication via mouse movements[C]// Proc of the 2004 ACM workshop on Visualization and data mining for computer security. 2004: 1-8

[11] Bours P, Fullu C J. A login system using mouse dynamics[C]// Proc of 5th International Conference on Intelligent Information Hiding and Multimedia Signal Processing. 2009: 1072-1077

[12] Schulz D A. Mouse curve biometrics [C]// Proc of 2006 Biometrics Symposium. 2006: 79-83

[13] Ahmed A A E, Traore I. A new biometric technology based on mouse dynamics[J]. IEEE Transactions on Dependable and Secure Computing, 2007, 4(3): 165-179

[14] Shen Chao, Cai Zhong-min, Guan Xiao-hong, et al. User authentication and monitoring based on mouse behavioral features[J]. Journal on Communications, 2010, 31(7): 68-75(in Chinese)
沈超,蔡忠闽,管晓宏,等. 基于鼠标行为特征的用户身份认证与监控[J]. 通信学报, 2010, 31(7): 68-75

(上接第 123 页)

[26] Jeh G, Widom J. SimRank: A measure of structural context similarity[C]//Proceedings of the ACM SIGKDD 2002. New York: ACM Press, 2002: 538-543

[27] Zhou T, Lv L, Zhang Y-C. Predicting missing links via local information[J]. European Physical Journal B, 2009, 71(4): 623-630

[28] Lv L, Jin C-H, Zhou T. Similarity index based on local paths for link prediction of complex networks[J]. Physical Review E-Statistical, Nonlinear, and Soft Matter Physics, 2009, 80(4): 046122

[29] Liu W-P, Lv L. Link Prediction Based on Local Random Walk [J]. European Physics Letter. , 2010, 89(5): 58007

[30] Rao Jun, Wu Bin, Dong Yu-xiao. Parallel Link Prediction in Complex Network Using MapReduce[J]. Journal of Software, 2012, 23(12): 3175-3186(in Chinese)
饶君,吴斌,东昱晓. MapReduce 环境下的并行复杂网络链路预测[J]. 软件学报, 2012, 23(12): 3175-3186

[31] Dong Yu-xiao, Ke Qing, Wu Bin. Link Prediction Based on Node Similarity[J]. Computer Science, 2011, 38(7): 162-164(in Chinese)
东昱晓,柯庆,吴斌. 基于节点相似性的链接预测[J]. 计算机科学, 2011, 38(7): 162-164

[32] http://www.linkprediction.org/index.php/link/resource/data

[33] Latora V, Marchiori M. Efficient behavior of small-world networks[J]. Phys. Rev. Lett. , 2001, 67: 198701-198704

[34] Watts D J, Strogatz S. Collective dynamics of 'small-world' networks[J]. Nature. , 1998, 393(6684): 440-442

[35] Newman M E J. Assortative mixing in networks[J]. Phys. Rev. Lett. , 2002, 89(20): 208701-208705

[36] Newman M E J. Scientific collaboration networks. I. network construction and fundamental results [J]. Physical Review E, 2001, 64: 0161311-061317

[37] Newman M E J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality [J]. Physical Review E, 2001, 64: 0161321-0161327

[38] Liu Ai-fen, Fu Chun-hua, Zhang Zeng-ping, et al. An Empirical Statistical Investigation on Chinese Mainland Movie Network [J]. Complex Systems and Complexity Science, 2007, 4(3): 10-16

[39] Robins G, Alexander M. Small worlds among interlocking directors: network structure and distance in bipartite graphs [J]. Computational & Mathematical organization Theory, 2004, 10(1): 69-94

[40] Battiston S, Catanzaro M. Statistical properties of corporate board and director networks [J]. European Physics Journal B, 2004, 38(2): 345-352

[41] Chen Wen-qin, Lu Jun-an, Liang Jia. Research in Disease-Gene Network Based on Bipartite Network Projection[J]. Complex Systems & Complexity Science, 2009, 6(1): 13-19

[42] Ergun G. Human sexual contact network as a bipartite graph [J]. Physica A, 2002, 308(1-4): 483-488

[43] Lambiotte R, Ausloos M. Uncovering collective listening habits and music genres in bipartite networks [J]. Physical Review E, 2005, 72(6): 066107

[44] Le Blond S, Guillaume J L, Latapy M. Clustering in P2P exchanges and consequences on performances[C]// Castro M, Renesse R, eds. Peer-to-Peer Systems IV. Berlin: Heidelberg, 2005, 193-204