• 复杂性科学 •

基于异质模体特征的社交网络链路预测

方祺娜，许小可*

(大连民族大学信息与通信工程学院 辽宁 大连 116600)

【摘要】现有链路预测方法大多是针对同质网络，没有考虑到真实网络多数是节点或连边性质具有差异的异质网络，无法充分利用不同类型节点或连边的拓扑结构信息。提出了一种基于异质模体特征的链路预测方法，将网络中的用户以性别差异作为节点类型划分，构建区分节点类型的异质模体特征进行异质网络中的链路预测。在此基础上，提出融合同质模体与异质模体特征的链路预测算法，相比现有预测方法在真实数据集上的性能，其 AUC 值提高了 17.0%～27.1%，Precision 值提高了 7.6%～20.1%。该方法可应用于在线社交网络中挖掘用户性别差异对人际交往的作用与影响，分析异质网络的演化动力学。

关 键 词 异质模体； 同质模体； 异质网络； 链路预测

中图分类号 TP391 文献标志码 A doi:10.12178/1001-0548.2021181

Link Prediction by Hetergeneous Motifs in Social Networks

FANG Qina and XU Xiaoke*

(College of Information and Communication Engineering, Dalian Minzu University Dalian Liaoning 116600)

Abstract Most of the existing link prediction methods are aimed at homogenous networks without considering that the real networks are heterogeneous networks with different node or edge properties. This kind of methods cannot make full use of the topological structure information of different types of nodes or edges. Under this circumstance, this paper proposes a link prediction method based on heterogeneous phantom features, which divides users in the network by gender differences as node types, and constructs heterogeneous phantom features distinguishing node types for link prediction in heterogeneous networks. On this basis, a link prediction algorithm that combines the characteristics of the homogeneous phantom and the heterogeneous phantom is proposed. Compared with the performance of the existing prediction method on the real data set, the AUC value is increased by 17.0%～23.1%, and the precision value is increased by 7.6%～20.1%. This method can be used in online social networks to explore the role and influence of user gender differences on interpersonal communication, and then to analyze the evolutionary dynamics of heterogeneous networks.

Key words heterogeneous motif; homogeneous motif; heterogeneous network; link prediction

信息时代，越来越多的人倾向于通过网络平台进行交流沟通[1]。互联网技术的快速发展使得社交网络的研究得到广泛关注[2]，如何对社交网络中复杂而庞大的用户关系进行预测和推荐是社交网络领域的研究热点，也是链路预测的重要应用方向[3]。链路预测能够揭示网络中用户之间的潜在关系[4]，挖掘社交用户的兴趣，为用户推荐朋友等，在社交服务中具有广泛应用[5]。

链路预测是网络挖掘中的一个基本问题[6]，也是复杂网络的研究热点。复杂网络根据结构可以分为同质网络和异质网络[7]。同质网络中的节点和连边为同一种类型，异质网络中的节点或连边为多种类型。目前大多数链路预测算法只考虑了网络的结构信息，没有考虑节点的属性[8]，已有社交网络链路预测问题的研究主要针对同质网络，针对异质网络的链路预测研究相对较少[9]。文献 [10] 提出基于异质网络表征学习的链路预测算法，通过元路径的随机游走实现网络表征学习进行异质网络链路预测。文献 [11] 根据元路径的质量权重建立预测模型，构建了一种基于元路径的链路预测方法。文

收稿日期：2021－07－08；修回日期：2021－11－22

基金项目：国家自然科学基金 (61773091, 62173065)；辽宁省“兴辽英才”计划项目 (XLYC1807106)；辽宁省自然科学基金 (2020-MZLH-22)

作者简介：方祺娜 (1996－)，女，主要从事社交网络数据挖掘和链路预测方面的研究.

*通信作者：许小可，E-mail：xuxiaoke@foxmail.com

献 [12] 通过挖掘有效、可用的元路径，提出基于图核的异质网络链路预测方法。虽然上述针对异质网络的链路预测方法取得了较好性能，但是它们主要采用元路径方法利用连边异质性进行链路预测，这类方法只考虑了网络中部分关系模式，因此还需要针对精细刻画多类型用户之间复杂的网络关系进行研究，如从网络的节点异质性角度挖掘拓扑结构特征进行精准预测。

在传统的同质网络链路预测研究中，最经典的方法是基于节点局部结构的相似性，如共同邻居、Adamic-Adar、资源分配指标 (resource allocation, RA)[13] 等。上述指标都是基于网络中的共同邻居特征，计算复杂度较低、准确率较高。然而如在以性别差异作为节点类型划分的异质网络中，由于只有异性节点之间有连边，同性节点之间无连边，网络中没有共同邻居节点，因此此类方法无法采用。文献 [13] 在共同邻居的基础上考虑三阶路径的因素，提出了预测准确率更高的局部路径 (local path, LP) 指标，Katz 指标在三阶路径的基础上进一步考虑了网络的所有路径。文献 [14] 提出了基于节点之间连接偏好的偏好连接相似性指标 (preferential attachment, PA)。文献 [15] 重点研究了二部图网络，提出了该类网络的 CAR 方法。与现有基于共同邻居的方法相比，该方法不仅基于网络中的公共节点以及共同邻居节点，同时引入共同邻居之间链接的组合。文献 [16] 基于 RA 指标研究了预测准确度更高的，针对三阶路径的 L3 方法。以上 5 种方法可以用于网络中缺少共同邻居节点的异质网络链路预测研究，作为进行比较的基准算法。

模体是指网络中出现频率较高的子图结构[17]，是一种重要的网络拓扑结构[18]。模体可用以研究拓扑结构中节点之间的交互模式，有助于理解复杂网络的局部结构和功能，是研究链路预测问题的重要方法。文献 [19] 最早提出利用模体结构进行有向网络链路预测分析，虽然基于模体特征进行链路预测的研究日益增多，但大多是在同质网络中进行分析。如文献 [20] 使用模体来描述刻画科学家合作的关系模式，并通过模体的组合对科学家合作网络进行预测。如果不区分节点类型来刻画网络的结构特征，就忽略了节点的类型差异，无法充分利用节点的异质信息。

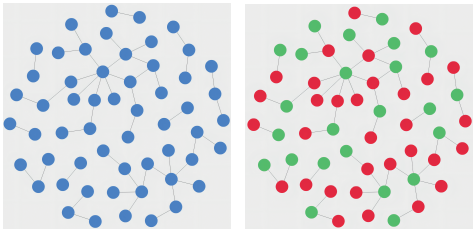
同质网络的链路预测研究往往不存在或者没有考虑节点的异质信息，存在一定的局限性。为了充分利用节点异质信息进行链路预测，本文提出基于异质模体特征的链路预测方法，将网络中不区分节点性别类型的模体结构定义为同质模体，区分节点性别类型的模体结构特征定义为异质模体，比较两种方法的预测性能差异和两种模体之间的关联性。为了结合不同模体特征的优势，本文还提出了融合同质模体与异质模体特征的链路预测算法。实验结果表明，相较于同质模体特征，基于异质模体特征的链路预测方法可以有效提升链路预测准确性，而融合同质和异质模体可以取得更好的预测效果。

1　问题描述及评价指标

1.1　问题描述

本文使用的社交网络为无向网络，形式为 $G(V,E)$，V、E分别是网络中的节点集合、连边集合。定义节点类型映射函数 $f: V \to A$，其中每个节点 $v \in V$ 都对应特定的类型 $f(v) \in A$；定义链接类型映射函数 $\gamma: E \to R$，其中每条链接 $e \in E$ 都对应特定的类型 $\gamma(e) \in R$。当 R 和 A 满足 $|A| > 1$ 或 $|R| > 1$ 时，即边的类型数或者节点的类型数大于 1，则该网络定义为异质网络，反之为同质网络。

本文将不区分用户类型的社交网络构建为同质网络，将用户类型区分为男性用户与女性用户的社交网络构建为异质网络。如图 1 所示，同质网络中的节点代表用户，异质网络中的浅色节点代表女性用户，深色节点代表男性用户。



a. 同质网络　　b. 异质网络

图 1　同质网络与异质网络

1.2　评价指标

1) 评价指标 AUC

AUC 作为衡量链路预测算法性能的一种重要指标，可以从整体上衡量算法的精确度[21]。AUC 指标可描述为如下形式：每次从测试集中随机选取一条存在的边，然后随机选取一条不存在的边，比较这两条边的相似度得分。如果存在边的分数大于不存在边的分数，就加 1 分；如果两条边的分数相等，就加 0.5 分。这样独立比较 n 次，如果有 n' 次存在边的分数值大于不存在边的分数值，有 n'' 次两条

边的分数值是相等的，则 AUC 值可以定义为：

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

通常，上述评分算法计算出的 AUC 值应该至少大于 0.5。AUC 的值越高，算法的精确度越高，但 AUC 的值最高不会超过 1。

2) 评价指标 Precision

Precision 作为衡量链路预测算法精确度的指标之一，主要从局部衡量预测的准确性。该指标关注的是预测值排序在前L个预测边中预测准确的比例。根据特征的分数值从大到小排序，如果有m条边是真实存在即预测准确的边，Precision 可以定义为：

$$\mathrm{Precision}=\frac{m}{L}$$

由该式可知，m越大则 Precision 值越高，预测越准确。

2 预测方法

2.1 基于相似性指标的预测方法

利用节点间的局域结构相似性是研究链路预测问题的一种重要方法，该方法的前提假设为节点间的相似性越大，它们之间存在链接的可能性就越大。在以往研究中，基于共同邻居相似性指标应用广泛、预测精度较高，但本文研究的异质社交网络数据由于只有不同类型的节点存在连边，故不存在共同邻居节点，因此无法基于共同邻居的相似性指标进行预测。本文主要使用局部路径指标 LP 与偏好连接相似性指标 PA、Katz、CAR 和 L3 作为链路预测的基准方法。LP 指标在考虑共同邻居的基础上考虑了三阶路径的因素，更全面考虑了节点的局域结构信息，可以有效提升预测精度；Katz 指标在三阶路径的基础上进一步考虑了网络的所有路径；PA 指标在网络存在“富者愈富”的连接偏好时性能显著，针对稀疏网络的预测性能也较好[22]；CAR 方法不仅考虑网络中的公共节点以及共同邻居节点，同时引入共同邻居节点之间链接的组合；L3 方法基于 RA 指标进一步提出三阶路径的预测方法，可以有效提升链路预测准确度。

1) 局部路径指标 (LP)：

$$S=\boldsymbol{A}^2+\alpha\boldsymbol{A}^3$$

式中，α为可调参数；$\boldsymbol{A}$表示网络的邻接矩阵，$(\boldsymbol{A})^n_{xy}$表示节点v_x和v_y之间长度为n的路径数。当$\alpha=0$时，LP 指标就等价于 CN 指标。

2) 偏好连接相似性 (PA)：

$$S_{xy}=k_xk_y$$

式中，k_n表示节点v_n的度，在网络中一条新边连接到节点v_n的概率正比于该节点的度k_n。在不考虑增长的网络中，新链接连接节点v_x和v_y的概率正比于两节点度k_xk_y的乘积。

3) 全局路径指标 (Katz)：

$$S_{xy}=\alpha\boldsymbol{A}_{xy}+\alpha^2\left(\boldsymbol{A}^2\right)_{xy}+\alpha^3\left(\boldsymbol{A}^3\right)_{xy}+\cdots$$

式中，$\alpha>0$为控制路径权重的可调参数，$(\boldsymbol{A})^n_{xy}$表示节点v_x和v_y之间长度为n的路径数。当参数α很小时，Katz 指标的预测结果接近于局部路径指标。

4) CAR 指标：

$$\mathrm{CAR}(x,y)=\mathrm{CN}(x,y)\mathrm{LCL}(x,y)$$

$$\mathrm{CN}(x,y)=|\Gamma(x)\cap\Gamma(y)|$$

$$\mathrm{LCL}(x,y)=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{|\gamma(z)|}{2}$$

式中，$\Gamma(x)$为节点x的邻居节点集合；$\Gamma(y)$为节点y的邻居节点集合；$\mathrm{CN}(x,y)$为节点x和节点y的三阶邻居数量；z为节点x和节点y的三阶邻居集合；$\gamma(z)$是节点z的局部社区度。

5) L3 指标：

$$\mathrm{L3}(x,y)=\sum_{u,v}\frac{a_{xu}a_{uv}a_{vy}}{\sqrt{k_uk_v}}$$

式中，k_u为节点u的度；a_{xu}代表节点x和节点u之间的相互作用。如果节点x和u之间存在相互作用，则$a_{xu}=1$，否则$a_{xu}=0$。

2.2 基于同质模体特征的预测方法

基于同质模体特征的链路预测方法主要是针对不考虑节点类型差异的同质网络，根据网络的拓扑结构，构建不区分节点类型的模体结构特征，将其定义为同质模体。由于本文数据为基于男女性别差异的异质网络数据，不考虑网络中的节点类型时，三节点模体和四节点模体结构只有表 1 的 5 种类型。

本文基于同质模体的预测方法共涉及 5 个模体特征，分别为 1 个三节点模体和 4 个四节点模体，代表了网络链接的 5 种关系模式。所有模体编号、图示和关系模式如表 1 所示，其中虚线表示待预测连边。

表 1　同质模体对应的关系模式

模体编号	图示	关系模式
T1		一位用户与两位无关系的用户其中的一位有社交关系，则可能与另一位用户有社交关系。
T2		一位用户与三位无关系的用户中的两位有社交关系，则可能与另一位用户有社交关系。
T3		两位无关系用户分别与两位有关系的用户具有关系，则其他两位用户可能有社交关系。
T4		两位有关系的用户，其中一位与另一位用户有关系，则另一位用户与其他用户也有关系。
T5		两位有社交关系的用户分别与两位无关系的用户有社交关系，则另两位用户可能有社交关系。

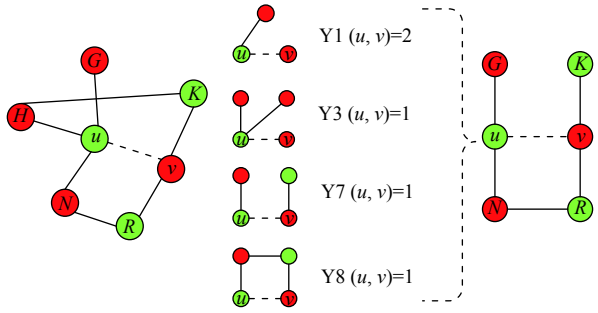
2.3　基于异质模体特征的预测方法

基于异质模特特征的链路预测方法主要针对异质网络，即网络中不只存在一种节点类型。根据异质网络的拓扑结构，构建区分节点类型的模体结构特征，将其定义为异质模体。本文主要基于男女性别进行节点类型区分，将节点分为男性节点与女性节点两种类型。在基于异质模体特征的预测方法中，三节点模体和四节点模体共涉及 8 种模体特征，分别为 2 个三节点模体和 6 个四节点模体，代表了社交网络中的 8 种关系模式。所有模体编号、图示和关系模式如表 2 所示，其中虚线表示待预测连边。

表 2　异质模体对应的关系模式

模体编号	图示	关系模式
Y1		一位男性用户与两位无关系的女性用户的其中的一位具有社交关系，则可能与另一位女性用户也有社交关系。
Y2		一位女性用户与两位无社交关系的男性用户其中一位具有社交关系，则可能与另一位男性也有社交关系。
Y3		一位女性用户与三位无社交关系的男性用户中的两位有关系，则可能与另一位男性有社交关系。
Y4		一位男性用户与三位无社交关系的女性用户中的两位有关系，则可能与另一位女性有社交关系。
Y5		两位有社交关系的男女用户，其中一位女性用户与一位男性用户有关系，则另外一位男性用户可能与其他女性有社交关系。
Y6		两位有社交关系的男女用户，其中男性用户与另外一位女性用户有社交关系，则另外一位女性用户可能与其他男性有社交关系。
Y7		两位无社交关系的男女用户分别与两位男女用户有社交关系，则这两位男女用户可能有社交关系。
Y8		除某两位男女用户无社交关系外，其余男女用户之间两两相互有关系，则无社交关系的男女用户可能有社交关系。

基于异质模体特征的社交网络关系预测主要提取训练集的模体特征，将每种预测边上的模体数量作为特征值，男性节点与女性节点之间是否有连边作为机器学习的分类标签，得到预测结果后使用 AUC 和 Precision 指标衡量预测性能。图 2 为基于异质模体特征的社交网络关系预测的具体过程。



a. 7 节点小网络图　　b. 用户关系预测的过程范例　　c. 组合结构

图 2　基于异质模体特征的关系预测

如图 2 所示，图 2a 为一个 7 节点的小型异质网络。本文数据为区分男女性别的异质网络数据，且只有男性节点与女性节点存在连边。图 2a 中节点u为男性节点，节点v为女性节点，边(u,v)为待预测连边，图 2b 中以异质模体特征 Y1、Y3、Y7、Y8 为例说明社交网络关系预测的主要过程，异质模体特征 Y1、Y3、Y7、Y8 的具体数量即为不同模体的特征值。模体特征 Y1 的计算方法为寻找节点 u 的邻居节点，且该邻居节点不是节点v的邻居。模体特征 Y7 的计算方法为寻找节点u和v各自的邻居节点，且该邻居节点不互为邻居。其他模体特征的计算方法以此类推，通过计算得出模体特征 Y1 的个数为 2，模体特征 Y3 的个数为 1，模体特征 Y7 的个数为 1，模体特征 Y8 的个数为 1。

在进行社交网络用户关系预测时，计算图 2c 中所涉及的 4 种模体在图 2a 小网络中的数量，并将得到的每种模体数量作为机器学习方法的输入，从而得到连边的相似度得分，继而进行网络的链路预测。

3　预测结果分析

3.1　实证数据说明

本文使用百度贴吧数据与性接触数据，分别构建同质网络与异质网络进行链路预测，网络具体信息如表 3 所示。

百度贴吧数据为百度贴吧恋爱吧用户评论数据，在该网络中，节点代表贴吧中的用户，依据性别划分为男性用户和女性用户，连边代表一名用户对另一名用户的发帖进行了评论或回复。本文将百

度恋爱吧男女之间的评论关系设定为具有线上社交关系，恋爱吧数据构建的网络，只使用男性节点与女性节点的社交关系构成连边。

性接触网络全称为基于性接触的经验时空网络 (empirical spatiotemporal network of sexual contacts[23])，该网络是一名男性用户与另一名女性用户进行性接触的线上沟通网络数据，节点代表性接触网络中的用户个体，分为男性用户与女性用户，连边代表一名男性用户与一名女性用户进行了线上的联络，即具有特殊社交关系。

表 3　实证网络信息说明

网络信息	节点/个	连边/条	男性节点/个	女性节点/个	男女节点比例
性接触网络	16730	50632	10106	6624	1.5
百度贴吧	2221	4990	1522	699	2.2

在进行链路预测实验时，对于每个实证网络数据，从正样本和负样本中分别随机选取 90% 的数据作为训练集E_T，选取剩余 10% 的正负样本数据作为测试集E_v，满足训练集与测试集正负样本比例 1:1。

3.2　基于模体特征链路预测

本文对所有单个模体特征 (5 个同质模体和 8 个异质模体) 和多个模体特征 (所有 5 个同质模体和所有 8 个异质模体) 进行链路预测，得到评价指标 AUC 与 Precision 的值。链路预测的结果如表 4 和表 5 所示，单个模体特征的最好预测性能和多模体特征的预测效果加粗标出。

表 4　基于同质模体特征的链路预测结果

模体编号	百度贴吧 AUC	性接触网络 AUC	百度贴吧 Precision	性接触网络 Precision
T1	0.522	0.521	0.551	0.547
T2	0.530	0.537	0.535	0.552
T3	**0.598**	**0.620**	**0.655**	**0.667**
T4	0.521	0.564	0.552	0.574
T5	0.527	0.637	0.541	0.645
多同质模体	**0.641**	**0.717**	**0.821**	**0.709**

由表 4 可以发现，使用单个同质模体特征进行链路预测时，模体特征 T3 的预测准确率和精确度最高。说明在社交网络中，如果两位无关系用户分别与两位其他用户具有社交关系，则其他两位用户有社交关系的可能性较大。本文综合多个同质模体特征进行预测，发现多同质模体特征的预测效果比单个同质模体特征的最好预测效果高 4.3%～16.6%，说明综合多种用户关系模式进行链路预测效果更好。

表 5　基于异质模体特征的链路预测结果

模体编号	百度贴吧 AUC	性接触网络 AUC	百度贴吧 Precision	性接触网络 Precision
Y1	0.567	0.576	0.621	0.589
Y2	0.518	0.547	0.516	0.501
Y3	0.546	0.545	0.558	0.587
Y4	0.516	0.541	0.524	0.614
Y5	0.590	0.568	0.658	0.587
Y6	0.522	0.534	0.518	0.507
Y7	**0.611**	**0.631**	**0.696**	**0.653**
Y8	0.592	0.621	0.601	0.711
多异质模体	**0.663**	**0.746**	**0.824**	**0.715**

由表 5 可以发现，使用单个异质模体特征进行链路预测时，模体特征 Y7 的预测准确率与精确度最高，说明在社交网络中，如果两位有关系的男女分别与两位无关系的男女有关系，则另外两位男女有关系的可能性越大。在 Y7 与 T3 的网络拓扑结构一致的情况下，异质模体特征的预测效果优于同质模体特征的预测效果。本文综合多个异质模体特征进行预测，发现多异质模体特征的预测效果比单个异质模体特征的最好预测效果高 5.2%～12.8%，说明综合多种男女用户关系模式进行链路预测效果更好。

除了比较链路预测的具体性能，本文还对 8 种异质模体特征进行了皮尔逊相关性分析，结果如图 3 所示。模体特征 Y1 和 Y3 具有较强相关性，Y2 和 Y4 也具有较强相关性，主要原因是 Y3 与 Y4 都是 Y1 与 Y2 的拓扑组合。Y7 与 Y8 也具有较强相关性，是因为这两个模体特征只关注待预测连边中两个节点的各自邻居节点之间的结构。Y1、Y2、Y5、Y7、Y8 可以视为一个相关性程度较高的集合，它们之间有较强的相关性，是因为它们的拓扑结构都是以 Y1 的拓扑结构为基础。

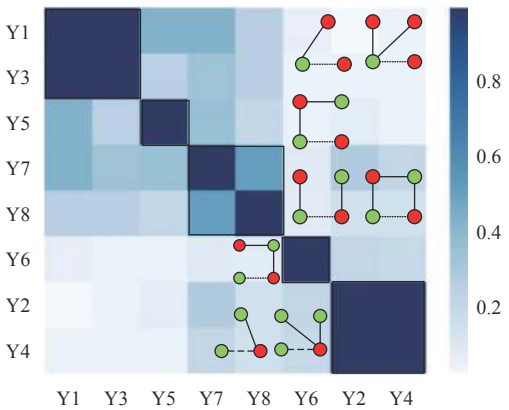
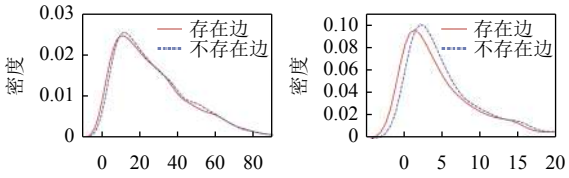


图 3　链路预测异质模体特征的相关性分析

3.3　同质模体特征与异质模体特征预测方法比较

为了比较同质模体特征与异质模体特征之间的差异，本文对两种模体结构存在边和不存在边的分布情况进行比较分析。百度贴吧数据中同质模体 T1 和异质模体 Y1 存在边和不存在边的分布差别如图 4 所示。其中实线和虚线分别代表网络中的存在边和不存在边的模体数量分布。研究发现，对于同质模体而言，存在边和不存在边有很大程度的重叠，重叠程度越大越不利于链路预测。对于异质模体，存在边和不存在边的重叠分布小于同质模体，说明相较于同质模体，使用异质模体进行链路预测的性能更好。



a. 同质模体边的分布　　b. 异质模体边的分布

图 4　同质模体和异质模体边的分布

本文采用基于同质模体特征方法与异质模体特征方法进行链路预测，在相同的网络拓扑结构下，同质模体和异质模体具有一定的相关性。图 5 分别为相同的网络拓扑结构下，同质模体与异质模体之间的关联性。其中节点代表用户个体，节点之间的连边代表用户之间的社交关系。深色节点代表用户性别为男性，浅色节点代表用户性别为女性。

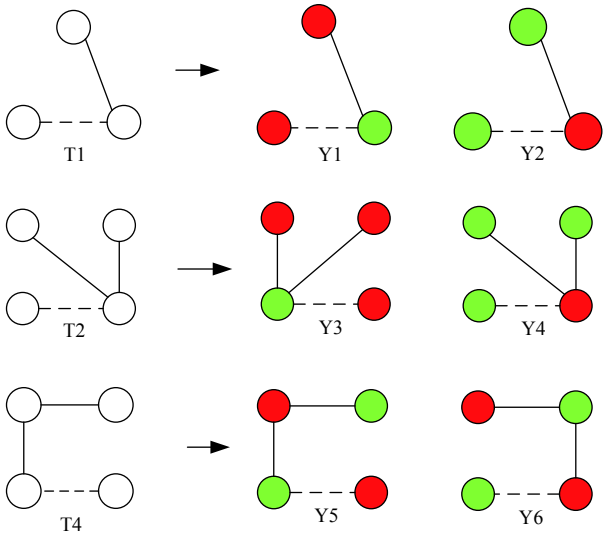


图 5　同质模体特征与异质模体特征结构差异

由图 5 可知，Y1、Y2 和 T1，Y3、Y4 和 T2，Y5、Y6 和 T4 分别具有相同的网络拓扑结构，为了探究相同网络拓扑结构下，考虑节点异质信息和不考虑节点异质信息的模体的链路预测效果，进行了基于单个异质模体特征、单个同质模体特征、同一网络拓扑结构下多异质模体特征的链路预测，结果如表 6 所示。

表 6　融合多个异质模体特征的链路预测结果

模体编号	百度贴吧 AUC	性接触网络 AUC	百度贴吧 Precision	性接触网络 Precision
T1	0.522	0.521	0.551	0.547
Y1	0.567	0.576	0.621	0.589
Y2	0.518	0.547	0.516	0.501
Y1+Y2	**0.578**	**0.611**	**0.711**	**0.697**
T2	0.530	0.537	0.535	0.552
Y3	0.546	0.545	0.558	0.587
Y4	0.516	0.541	0.524	0.614
Y3+Y4	**0.569**	**0.576**	**0.593**	**0.606**
T4	0.521	0.564	0.552	0.574
Y5	0.590	0.568	0.658	0.587
Y6	0.522	0.534	0.518	0.507
Y5+Y6	**0.601**	**0.612**	**0.647**	**0.664**

通过表 6 可以发现，在两个实证网络数据中，融合多个异质模体特征的 AUC 和 Precision 值均高于单个异质模体特征和同质模体特征。结果表明在相同的网络拓扑结构下，融合所有区分节点异质信息的异质模体特征，其链路预测准确性高于单个异质模体特征以及不考虑节点异质信息的同质模体特征。这是由于异质模体考虑了网络中节点的异质信息，更全面准确地刻画了网络结构。

3.4　融合同质模体和异质模体特征的链路预测

以往关于链路预测的研究中，研究人员提出的基于网络结构相似性的方法大多只关注其中一种网络结构，即一种模体结构。在应用于社交网络的链路预测算法中，往往也只研究了一种社交用户之间的关系模式，忽略了社交用户之间多种关系模式的组合。因此本文通过特征拼接的方式融合多种同质模体和异质模体结构进行链路预测，旨在结合不同模体特征的优势，分析多模体结构即多关系模式对链路预测准确性的影响，并将多模体结构的预测结果与单模体结构的预测结果进行比较。

在链路预测问题中，将所有同质模体特征与所有异质模体特征进行融合，链路预测的结果如表 7 所示，发现融合多同质模体和异质模体特征的链路预测准确率高于只使用多异质模体特征的链路预测准确率。说明相较于只使用多异质模体进行链路预测，融合同质模体特征对提升链路预测准确性具有

一定的积极作用。本文还将所有同质模体特征、所有异质模体特征、融合所有异质模体和同质模体特征与 LP、Katz、PA 和 CAR 和 L3 进行了对比，结果如表 7 所示，其中最好的预测效果已加粗标出。

表 7　5 类方法的链路预测结果

网络	评价指标	LP	PA	Katz	CAR	L3	同质模体	异质模体	同质模体+异质模体
百度贴吧	AUC	0.532	0.541	0.543	0.652	0.560	0.641	0.663	**0.702**
	Precision	0.567	0.585	0.591	0.661	0.544	0.717	0.746	**0.768**
性接触网络	AUC	0.638	0.598	0.672	0.656	0.571	0.821	0.824	**0.869**
	Precision	0.653	0.601	0.667	0.680	0.550	0.709	0.715	**0.729**

由表 7 中数据可知，融合多异质模体和同质模体特征的链路预测算法准确率最高，其 AUC 比 LP、PA、Katz 方法最多提升了 27.1%，精确度最多提高了 20.1%，该方法也优于 CAR 和 L3 方法的精确度。这是因为相比 CAR 和 L3 方法，本文提出的基于多同质模体和多异质模体的链路预测方法考虑了更多网络结构的非局域信息。因此，在社交网络中融合多同质和异质模体特征进行链路预测能够有效提高预测的准确性。

尽管 CN、LP 等局部相似性指标可使用坚实的理论和实证依据进行解释，如社会学中的同质性原理，即两个相似的节点更大概率产生连边[16]。但最新研究发现，并不存在某一类局域指标可在所有实证网络中都取得最佳预测性能，有些网络是基于二阶路径的相似性指标表现更好，而另一些是三阶路径指标取得更好性能。本文以特殊的异质社交网络为研究对象，这类网络的突出特点是局域性指标失效而只能依靠刻画结构非局域性的模体结构进行链路预测，因此对于研究其他网络的非局域性指标具有一定的借鉴作用，同时考虑到节点角色的异质性也有利于将此类方法应用于二部分图中[24]。

由于本文数据为实证网络数据，每位用户可能存在造假的动机和现象。为了验证当节点的男女信息存在噪音情况下算法结果的稳定性，本文以百度贴吧数据为例，进行男女节点性别互换。随机选取实证数据中 30%、40%、50%、60% 的男女节点进行性别互换，互换后的链路预测结果如图 6 所示。

由图 6 可知，虽然对实证数据中的男女性别进行了一定比例的置乱，但实验结果表明依旧是多同质模体与异质模体的链路预测算法准确性最高，其次是多异质模体，均高于同质模体的准确性。该结果与上文的实验结果一致，因此本文算法具有一定的通用性和稳定性。

在融合所有同质模体和异质模体特征的链路预测中，本文还对 8 种异质模体和 5 种同质模体进行皮尔逊相关性分析，结果如图 7 所示。

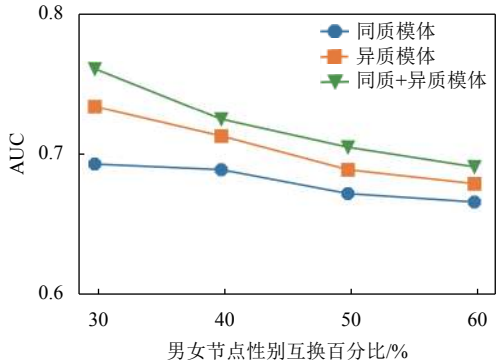


图 6　男女性别互换的链路预测结果

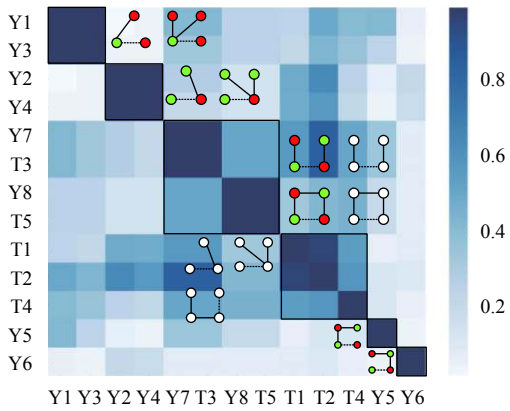


图 7　链路预测同质模体与异质模体特征的相关性分析

由图 7 可看出，异质模体 Y1 和 Y3，Y2 和 Y4 具有较强相关性，原因是模体特征 Y3 与 Y4 分别是模体特征 Y1 与 Y2 拓扑结构的组合。异质模体特征 Y7 和 Y8 与同质模体特征 T3 和 T5 具有较强相关性，原因是这 4 种模体的网络拓扑结构较为接近，都是以四节点方形拓扑结构为基础进行模体的构建。同质模体 T1、T2 和 T4 相关性也较强，这是由于 3 种模体结构均为同质模体且拓扑结构都是以 T1 的拓扑结构为基础。

4　结 束 语

本文研究性接触网络与百度贴吧恋爱吧两种特殊类型网络，为了更精准地刻画网络结构以及充分

利用节点的异质信息，本文提出了基于异质模体的链路预测方法，验证了异质模体数量与链路预测准确率的相关性，构建异质模体特征进行关系预测。在此基础上，提出融合多种同质和异质模体特征进行社交网络链路预测方法。结果表明，基于异质模体的预测方法可以有效提升链路预测准确性，融合多异质和同质模体特征的预测效果更为显著。本研究有助于对社交网络的用户关系进行预测和推荐，在用户行为分析、推荐系统等方面具有广阔的应用前景。后续研究将在异质模体特征的基础上引入朴素贝叶斯算法与角色函数，对异质网络中的信息进行更加充分的利用。

周涛教授对本文研究工作给予了一些指导和帮助，在此表示感谢。

参　考　文　献

[1] SHABAZ M, GARG U. Shabaz-Urvashi link prediction (SULP): A novel approach to predict future friends in a social network[J]. Journal of Creative Communications, 2021, 16(5439): 27-44.

[2] WANG G H, WANG Y F, LI J M, et al. A multidimensional network link prediction algorithm and its application for predicting social relationships[J]. Journal of Computational Science, 2021, 53(5): 101358.

[3] SONGMUANG P, SIRISUP C, SUEBSRIWICHAI A. Missing link prediction using non-overlapped features and multiple sources of social networks[J]. Information, 2021, 12(5): 214-214.

[4] YAN R, LI Y, LI D, et al. SSDBA: The stretch shrink distance based algorithm for link prediction in social networks[J]. Frontiers of Computer Science, 2020, 15(1): 151301.

[5] DEBASIS D. Positive and negative link prediction algorithm based on sentiment analysis in large social networks[J]. Wireless Personal Communications, 2018, 102(3): 2183-2198.

[6] LYU L, ZHOU T. Link prediction in complex networks: A survey[J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150-1170.

[7] SHAO Y B, LIU C. H2Rec: Homogeneous and heterogeneous network embedding fusion for social recommendation[J]. International Journal of Computational Intelligence Systems, 2021, 14(1): 1303-1314.

[8] ZHOU T. Progresses and challenges in link prediction[J]. iScience, 2021, 24(11): 103217.

[9] BERAHMAND, KAMAL, et al. A modified deepwalk method for link prediction in attributed social network[J]. Computing, 2021, 103(4): 2227-2249.

[10] 蒋宗礼, 管戈. 基于异质网络表征学习的链路预测算法[J]. 现代计算机, 2020, 17(6): 29-33,37.
JIANG Z L, GUAN G. Link prediction algorithm based on heterogeneous network representation learning[J]. Modern Computer, 2020, 17(6): 29-33,37.

[11] 黄立威, 李德毅, 马于涛, 等. 一种基于元路径的异质信息网络链路预测模型[J]. 计算机学报, 2014, 37(4): 848-858.
HUANG L W, LI D Y, MA Y T, et al. A meta path-based link prediction model for heterogeneous information networks[J]. Chinese Journal of Computers, 2014, 37(4): 848-858.

[12] 赵妍, 赵书良, 马秋微. 基于图核的异质信息网络链路预测方法[J/OL]. 计算机应用研究, 2021, 6(4): 11-18.
ZHAO Y, ZHAO S L, MA Q W. Link prediction method of heterogeneous information network based on graph kernel application[J/OL]. Research of Computers, 2021, 6(4): 11-18.

[13] ZHOU T, LYU L Y, ZHANG Y C. Predicting missing links via local information[J]. The European Physical Journal B, 2009, 71(4): 623-630.

[14] SHAN Z. Link prediction based on local information considering preferential attachment[J]. Physica A: Statistical Mechanics and its Applications, 2016, 443(6): 537-542.

[15] DAMINELLI S, THOMASM, DURÁN C, et al. Common neighbours and the local-community-paradigm or topological link prediction in bipartite networks[J]. New Journal of Physics, 2015, 17(11): 113037.

[16] 李艳丽, 周涛. 链路预测中的局部相似性指标[J]. 电子科技大学学报, 2021, 50(3): 422-427.
LI Y L, ZHOU T. Local similarity indices in link prediction[J]. Journal of University of Electronic Science and Technology of China, 2021, 50(3): 422-427.

[17] PARK Y, LEE M, SON S. Motif dynamics in signed directional complex networks[J]. Journal of the Korean Physical Society, 2021, 78(6): 535-541.

[18] MILO R, SHEN-ORR S, ITZKOVITZ S, et al. Network motifs: Simple building blocks of complex networks[J]. Science, 2002, 298(5594): 824-827.

[19] ZHANG Q M, LYU L, WANG W Q, et al. Potential theory for directed networks[J]. PloS One, 2013, 8(2): e55437.

[20] 曹红艳, 许小可, 许爽. 基于多模体特征的科学家合作预测[J]. 电子科技大学学报, 2020, 49(5): 766-773.
CAO H Y, XU X K, XU S. Predicting scientist cooperation based on multiple motif features[J]. Journal of University of Electronic Science and Technology of China, 2020, 49(5): 766-773.

[21] 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.
LYU L Y. Link prediction on complex networks[J]. Journal of University of Electronic Science and Technology of China, 2010, 39(5): 651-661.

[22] FURQAN A, HAJI G, ISHTIAQ M, et al. Link prediction using node information on local paths[J]. Physica A: Statistical Mechanics and its Applications, 2020, 557(1): 124980.

[23] ROCHA L E C, LILJEROS F, HOLME P. Simulated epidemics in an empirical spatiotemporal network of 50, 185 sexual contacts[J]. PLoS Computational Biology, 2011, 7(3): e1001109.

[24] ZHOU T, LI Y L, WANG G. Experimental analyses on 2-hop-based and 3-hop-based link prediction algorithms[J]. Physica A, 2021, 564: 125532.

编　辑　蒋　晓