复杂网络链路预测

吕琳媛

(瑞士弗里堡大学物理系　瑞士 弗里堡　CH-1700)

【摘要】网络中的链路预测是指如何通过已知的网络结构等信息预测网络中尚未产生连边的两个节点之间产生连接的可能性。预测那些已经存在但尚未被我们发现的连接实际上是一种数据挖掘的过程，而对于未来可能产生的连边的预测则与网络的演化相关。传统的方法是基于马尔科夫链或者机器学习的，往往考虑节点的属性特征。这类方法虽然能够得到较高的预测精度，但是由于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。另一类方法是基于网络结构的最大似然估计，这类方也有计算复杂度高的问题。相比上述两种方法，基于网络结构相似性的方法更加简单。通过在多个实际网络中的实验发现，基于相似性的方法能够得到很好的预测效果，并且网络的拓扑结构性质能够帮助选择合适的相似性指标。本文综述并比较了若干有代表性的链路预测方法，展望了若干重要的开放性问题，可望为相关学者提供借鉴。

关　键　词　链路预测;　复杂网络;　相似性指标;　最大似然估计;　概率模型

中图分类号　TP391　　文献标识码　A

Link Prediction on Complex Networks

Linyuan Lü

（Department of Physics, University of Fribourg, Fribourg CH-1700, Switzerland)

Abstract： Link prediction aims at estimating the likelihood of the existence of links between nodes. The predicting of existed yet unknown links is similar to the data mining process, while the predicting of future links relates to the network evolution. The traditional methods are based on Markov Chains and machine learning which usually involve the node attributes information. Although these methods can give good prediction, the high computational complexity limits their applications in large-scale systems. The approaches based on maximum likelihood approximation also suffer this shortcoming. Another group of methods is based on the node similarity that is defined base solely on the network structure. Extensive experiments on many real networks show that the similarity-based methods can give good prediction while with lower computational complexity comparing with the above two kinds of methods. This article introduces and compares many representative link prediction methods, and outlines some important open problems, which may be valuable for related research domains.

Key words: Complex Networks; Link Prediction; Maximum likelihood approximation; Probabilistic model; Similarity Index

1.　前言

网络中的链路预测(Link Prediction)是指如何通过已知的网络节点以及网络结构等信息预测网络中尚未产生连边的两个节点之间产生连接的可能性[1]。这种预测既包含了对未知链接（exist yet unknown links）的预测，也包含了对未来链接（future links)的预测。链路预测作为数据挖掘领域的研究方向之一在计算机领域已有较深入的研究。他们的研究思路和方法主要基于马尔科夫链和机器学习。Sarukkai[2]应用马尔科夫链进行网络的链路预测和路径分析。之后 Zhu 等人[3]将基于马尔科夫链的预测方法扩展到了自适应性网站（adaptive web sites）的预测中。此外，Popescul 和 Ungar[4]提出一个回归模型在

收稿日期：2010－07－18；修回日期：2010－07－22

基金项目：瑞士国家科学基金项目(200020-121848), 国家自然科学基金项目(60973069)

作者简介：吕琳媛(1984－)，女，博士研究生，主要从事信息物理，包括链路预测，推荐算法以及网络节点排序等方面的研究.

文献引用网络中预测科学文献的引用关系。他们的方法不仅用到了引文网络的信息还有作者信息、期刊信息以及文章内容等外部信息。应用节点属性的预测方法还有很多，例如O'Madadhain等人[5]利用网络的拓扑结构信息以及节点的属性建立了一个局部的条件概率模型来进行预测。Lin[6]基于节点的属性定义了节点间的相似性，可以直接用来进行链路预测。虽然应用节点属性等外部信息的确可以得到很好的预测效果，但是很多情况下这些信息的获得是非常困难的，甚至是不可能的。比如很多在线系统的用户信息都是保密的。另外即使获得了节点的属性信息也很难保证信息的可靠性，即这些属性是否反映了节点的真实情况，例如在线社交网络中很多用户的注册信息都是虚假的。更进一步，在能够得到节点属性的精确信息的情况下，如何鉴别出哪些信息对网络的链路预测是有用的，哪些信息是没用的仍然是个问题。

最近几年，基于网络结构的链路预测方法受到越来越多的关注。相比节点的属性信息而言，网络的结构更容易获得，也更加可靠。同时这类方法对于结构相似的网络具有普适性，从而避免了对不同网络需要机器学习获得一些特定的参数组合。Liben-Nowell 和Kleinberg[7]提出了基于网络拓扑结构的相似性定义方法，并将这些指标分为基于节点和基于路径的两类，并分析了若干指标对社会合作网络中链路预测的效果。另外一类链路预测方法是基于网络结构的最大似然估计。Clauset, Moore 和 Newman 于 08 年发表在《自然》上的论文提出了一种利用网络的层次结构进行链路预测的方法，并在具有明显层次结构的网络中表现很好[8]。此外 09 年底 Guimera 和 Sales-Pardo 在《美国科学院院刊》(PNAS)上发表了一篇利用随机分块模型[9]预测网络缺失边和错误边的链路预测方法[10]。值得一提的是这篇文章第一次提到网络错误链边（spurious links）的概念，即在网络已知的链接中很可能存在一些错误的链接，比如我们对蛋白质相互作用关系的错误认知。

链路预测问题受到来自不同领域拥有不同背景的科学家的广泛关注，首先是因其重大的实际应用价值。在生物领域研究中，例如蛋白质相互作用网络和新陈代谢网络，节点之间是否存在链接，或者说是否存在相互作用关系，是需要通过大量实验结果进行推断的。我们已知的实验结果仅仅揭示了巨大网络的冰山一角。仅以蛋白质相互作用网络为例，酵母菌蛋白质之间 80%的相互作用不为我们所知[11]，而对于人类自身，我们知道的仅有可怜的 0.3%[12,13]。由于揭示这类网络中隐而未现的链接需要耗费高额的实验成本。那么如果能够事先在已知网络结构的基础上设计出足够精确的链路预测算法，再利用预测的结果指导试验，就有可能提高实验的成功率从而降低试验成本，并加快揭开这类网络真实面目的步伐。实际上，社会网络分析中也会遇到数据不全的问题，这时候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[14,15]。除了帮助分析数据缺失的网络，链路预测算法还可以用于分析演化网络。举例来说，近几年在线社交网络发展非常迅速[16]，链路预测可以基于当前的网络结构去预测哪些现在尚未结交的用户“应该是朋友”，并将此结果作为“朋友推荐”发送给用户。如果预测足够准确，显然有助于提高相关网站在用户心目中的地位，从而提高用户对该网站的忠诚度。另外，链路预测的思想和方法，还可以用于在已知部分节点类型的网络(partially labeled networks)中预测未标签节点的类型——这可以用于判断一篇学术论文的类型[17]或者判断一个手机用户是否产生了切换运营商(例如从移动到联通)的念头[18]。另外 Guimera 和 Sales-Pardo 所提出的对网络中的错误链接的预测[10]，对于网络重组和结构功能优化也有重要的应用价值。例如在很多构建生物网络的实验中存在暧昧不清甚至自相矛盾的数据[19]，我们就有可能应用链路预测的方法对其进行纠正。

链路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，特别是对一些相关领域理论方面的推动和贡献。近年来，随着网络科学的快速发展，其理论上的成果为链路预测搭建了一个研究的平台，使得链路预测的研究与网络的结构与演化紧密联系

起来。因此，对于预测的结果更能够从理论的角度进行解释。与此同时，链路预测的研究也可以从理论上帮助我们认识复杂网络演化的机制。针对同一个或者同一类网络，很多模型都提供了可能的网络演化机制[20,21]。由于刻画网络结构特征的统计量非常多，很难比较不同的机制孰优孰劣。链路预测机制有望为演化网络提供一个简单统一且较为公平的比较平台，从而大大推动复杂网络演化模型的理论研究。另外，如何刻画网络中节点的相似性也是一个重大的理论问题[22]，这个问题和网络聚类等应用息息相关[23]。类似地，相似性的度量指标数不胜数，只有能够快速准确地评估某种相似性定义是否能够很好刻画一个给定网络节点间的关系，才能进一步研究网络特征对相似性指标选择的影响。在这个方面，链路预测可以起到核心技术的作用。链路预测问题本身也带来了有趣且有重要价值的理论问题，也就是通过构造网络系综并藉此利用最大似然估计的方法进行链路预测的可能性和可行性研究。这方面的研究对于链路预测本身以及复杂网络研究的理论基础的建立和完善，可以起到推动和借鉴的作用。

2. 问题描述与评价方法

定义 G（V，E）为一个无向网络，其中 V 为节点集合，E 为边集合。网络总的节点数为 N，边数为 M。此网络共有 $N(N-1)/2$ 个节点对，即全集 U。给定一种链路预测的方法，对每对没有连边的节点对 x，y（$\in U \setminus E$）赋予一个分数值 S_{xy}。然后将所有未连接的节点对按照该分数值从大到小排序，排在最前面的节点对出现连边的概率最大。

为了测试算法的准确性，将已知的连边 E 分为两部分，训练集 E^T 和测试集 E^P。在计算分数值的时候只能使用测试集中的信息。显然，$E = E^T \bigcup E^P$，且 $E^T \bigcap E^P = \varnothing$。在此，将属于 U 但不属于 E 的边定义为不存在的边。衡量链路预测算法精确度的指标有三种 AUC，Precision 和 Ranking Score。他们对预测精确度衡量的侧重点不同：AUC（area under the receiver operating characteristic curve）是从整体上衡量算法的精确度[24]， Precision 只考虑排在前 L 位的边是否预测准确[25]，而 Ranking Score 更多考虑了所预测的边的排序[26]。

AUC 可以理解为在测试集中的边的分数值比随机选择的一个不存在的边的分数值高的概率。也就是说每次随机从测试集中选取一条边与随机选择的不存在的边进行比较，如果测试集中的边分数值大于不存在的边的分数，那么就加一分，如果两个分数值相等就加 0.5 分。这样独立的比较 n 次，如果有 n' 次测试集中的边分数值大于不存在的边分数，有 n'' 次两分数值相等，那么 AUC 定义为：

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

显然，如果所有分数都是随机产生的，那么 AUC=0.5。因此 AUC 大于 0.5 的程度衡量了算法在多大程度上比随机选择的方法精确。

Precision 定义为在前 L 个预测边中有几个预测准确的比例。如果有 m 个预测准确，即排在前 L 的边中有 m 个在测试集中，那么 Precision 定义为：

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

显然，Precision 越大预测越准确。如果两个算法 AUC 相同，而算法 1 的 Precision 大于算法 2，那么说明算法 1 更好，因为他倾向于把真正连边的节点对排在前面。

Ranking Score 主要考虑测试集中的边在最终排序中的位置。令 $H = U - E^T$ 为未知边

的集合（相当于测试集中的边和不存在的边的集合），r_i表示未知边$i \in E^P$在排序中的排名。那么这条未知边的 Ranking Score 值为$RS_i = r_i / |H|$，遍历所有在测试集中的边得到系统的 Ranking Score 值为

$$RS = \frac{1}{|E^P|}\sum_{i \in E^P} RS_i = \frac{1}{|E^P|}\sum_{i \in E^P} \frac{r_i}{|H|}$$

3. 基于相似性的链路预测

应用节点间的相似性进行链路预测的一个重要前提假设就是两个节点之间相似性(或者相近性)越大，它们之间存在链接的可能性就越大。注意这里所指的相似性并非一般意义上的相似性，而是指一种接近程度（Proximity）。刻画节点的相似性有很多种方法。最简单直接的就是利用节点的属性，例如，如果两个人具有相同的年龄，性别，职业，兴趣，等等，我们说他们俩很相似。利用节点属性的相似性进行链路预测的前提就是网络中的边本身代表着相似。另外一类相似性的定义完全基于网络的结构信息，我们称之为结构相似性。基于结构相似性的链路预测精度的高低取决于这种结构相似性的定义是否能够很好的抓住目标网络的结构特征。例如基于共同邻居的相似性指标，即两个节点如果有更多的共同邻居就会更可能连边，在集聚系数较高的网络中表现非常好，有时甚至超过一些更复杂的算法。然而对于集聚系数较低的网络如路由器网络或电力网络等，预测精度就差很多。

3.1 基于局部信息的相似性指标

基于局部信息的最简单的相似性指标是共同邻居（Common Neighbors），也就是说两个节点如果有更多的共同邻居那么他们更倾向于连边。在共同邻居的基础上考虑两端节点度的影响从不同的角度以不同的方式又产生 6 种相似性指标，分别是 Salton 指标[27]（也叫做余弦相似性），Jaccard 指标[28]，Sorenson 指标[29]，大度节点有利指标（Hub Promoted Index）[30]，大度节点不利指标（Hub Depressed Index），LHN-Ⅰ指标[22]（由 Leicht，Holme 和 Newman 提出而得名）。我们称这一类指标为基于共同邻居的相似性。

另一个只考虑节点度的相似性为优先连接指标（Preferential Attachment）。应用优先连接的方法可以产生无标度的网络结构，在这种网络中，一条即将加入的新边连接到节点 x 的概率正比于节点 x 的度 $k(x)$[31]，因此新边连接节点 x 和 y 的概率就正比于两节点度的乘积。此算法的复杂度较其它算法来说最低，因为需要的信息量最少。

如果考虑两节点共同邻居的度信息，有 Adamic-Adar 指标[32]。他的思想是度小的共同邻居节点的贡献大于度大的共同邻居节点。因此根据共同邻居节点的度为每个节点赋予一个权重值，该权重等于该节点的度的对数分之一，即 $1/\log k$。

周涛、吕琳媛和张翼成从网络资源分配（Resource Allocation）的角度提出一种新的指标，简称 RA [33]。考虑网络中没有直接相连的两个节点 x 和 y，从 x 可以传递一些资源到 y，而在此过程中他们的共同邻居就成为传递的媒介。假设每个媒介都有一单位的资源并且将平均分配传给他的邻居，则 y 可以接收到的资源数就可定义为节点 x 和 y 的相似度。RA 和 AA 指标最大的区别就在于赋予共同邻居节点权重的方式不同，前者以 $1/k$ 的形式递减，后者以 $1/\log k$ 的形式。可见当网络的平均度较小的时候 RA 和 AA 差别不大，但是当平均度较大的时候就有很大区别了。

表 1 总结了以上 10 种基于局部信息的相似性指标的定义公式。对于网络中的节点 x，定义它的邻居为$\Gamma(x)$，$k(x) = |\Gamma(x)|$为节点 x 的度。

名称	定义	名称	定义
共同邻居（CN）	$s_{xy}=\lvert\Gamma(x)\cap\Gamma(y)\rvert$	大度节点不利指标（HDI）	$s_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\max\{k(x),k(y)\}}$
Salton 指标[27]	$s_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\sqrt{k(x)\times k(y)}}$	LHN-I 指标[22]	$s_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{k(x)\times k(y)}$
Jaccard 指标[28]	$s_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$	优先链接指标（PA）[31]	$s_{xy}=k(x)\times k(y)$
Sorenson 指标[29]	$s_{xy}=\frac{2\lvert\Gamma(x)\cap\Gamma(y)\rvert}{k(x)+k(y)}$	Adamic-Adar 指标（AA）[32]	$s_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{\log k(z)}$
大度节点有利指标（HPI）[30]	$s_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\min\{k(x),k(y)\}}$	资源分配指标（RA）[33]	$s_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{k(z)}$

表 1 十种基于节点局部信息的相似性指标

周涛、吕琳媛和张翼成将上述 10 种基于节点局部信息的相似性指标在六个实际网络中进行实验，并比较其预测精确度[33]。这六个网络分别为：蛋白质相互作用网络（PPI），科学家合作网络（NS），美国电力网络（Grid），政治博客网络（PB），路由器网络（INT）以及美国航空网络（USAir）。它们的统计性质见表 2。其中 N，M 分别表示网络的节点数和边数，Nc 为网络的最大联通集团，例如 2375/92 表示 PPI 网络中有 92 个连通集团，最大连通集包含 2375 个节点，e 为网络的效率，C 为网络集聚系数，r 为同配系数，H 为网络度的异质性。

Networks	**N**	**M**	**Nc**	**e**	**C**	**r**	**H**
PPI	2617	11855	2375/92	0.180	0.387	0.461	3.73
NS	1461	2742	379/268	0.016	0.878	0.462	1.85
Grid	4941	6594	4941/1	0.063	0.107	0.003	1.45
PB	1224	19090	1222/2	0.397	0.361	-0.079	3.13
INT	5022	6258	5022/1	0.167	0.033	-0.138	5.05
USAir	332	2126	332/1	0.406	0.749	-0.208	3.46

表 2 六个实验网络的拓扑性质

预测结果如表 3 所示。所有结果均以 AUC 为预测精度评价指标。可见在这十种算法中 RA 表现最好，其次是 CN，再其次是 AA 指标。总的来说 PA 表现最差，特别是在电力网络和路由器网络中预测精度还不到 0.5，这意味着 PA 算法在这两个网络中预测精度还不如完全随机的预测好。

Index	PPI	NS	Grid	PB	INT	USAir
CN	0.889	0.933	0.590	0.925	0.559	0.937
Salton	0.869	0.911	0.585	0.874	0.552	0.898

Jaccard	0.888	0.933	0.590	0.882	0.559	0.901
Sorensen	0.888	0.933	0.290	0.881	0.559	0.902
HPI	0.868	0.911	0.585	0.852	0.552	0.857
HDI	0.888	0.933	0.590	0.877	0.559	0.895
LHN-I	0.866	0.911	0.585	0.772	0.552	0.758
PA	0.828	0.623	0.446	0.907	0.464	0.886
AA	0.888	0.932	0.590	0.922	0.559	0.925
RA	0.890	0.933	0.590	0.931	0.559	0.955

表 3 十种基于节点局部信息的相似性在六个网络链路预测中的精度比较

3.2 基于路径的相似性指标

基于路径的相似性指标有三个，分别是局部路径性指标（LP）[34]，Katz 指标[35]，和 LHN-II [22]指标（与 LHN-I 在同一篇文章中提出）。

（1）局部路径指标（Local Path）是在共同邻居指标的基础上考虑了三阶邻居的贡献。其定义为 $S = A^2 + \alpha A^3$，其中 α 为可调节参数，用来控制三阶路径的作用大小，显然当 $\alpha = 0$ 时 LP 指标就等于 CN。A 为网络的邻接矩阵。注意 $(A^n)_{xy}$ 表示节点 x 和 y 之间长度为 n 的路径数。

（2）Katz 指标考虑的是所有的路径数，且对于短路径赋予较大的权重，而长路径赋予较小的权重。它定义为 $S = \beta A + \beta^2 A^2 + \beta^3 A^3 \cdots = (I - \beta A)^{-1} - I$，其中 β 为权重衰减因子，为了保证数列的收敛性 β 的取值须小于邻接矩阵 A 最大特征值的倒数。

（3）LHN-II 指标和 Katz 参数类似也是考虑所有路径。所不同的是 LHN-II 中每一项不再是 $(A^n)_{xy}$，而变为 $(A^n)_{xy} / E[(A^n)_{xy}]$，其中 $E[(A^n)_{xy}] = \frac{k_x k_y}{M} \lambda_1^{n-1}$，为节点 x 和 y 之间长度为 n 的路径数的期望值。整理后得到 LHN-II 最终表达式为 $S = M\lambda_1 D^{-1}(I - \frac{\phi A}{\lambda_1})^{-1} D^{-1}$，其中 λ_1 为 A 的最大特征值，ϕ 为参数取值小于 1（具体推导过程参见文献[22]）。

应用上述三种基于路径的相似性指标进行链路预测，结果总结于表 4，分别用 AUC 和 Precision（L=100）进行评价。LP 的结果是在最优参数 α 时得到的。LP*的结果是在固定参数 $\alpha = 0.01$ 时得到的。由于美国航空网络特殊的层次结构，在 USAir 网络中设定 $\alpha = -0.01$ [33]。从表中可以看出应用 AUC 作为评价指标的时候基于全局信息的 Katz 指标表现最好，特别是在电力网和 Internet 路由器网络中 AUC 可达到 0.95 以上。其次局部路径算法表现也不错，比如在 PPI 和 PB 网络中可以达到和 Katz 指标差不多好的预测精度。甚至在 PB 和 USAir 网络中表现比 Katz 指标还好。其原因在于 PB 和 USAir 网络的平均最短距离很小，因此基于 3 阶路径的 LP 指标比基于全部路径的 Katz 指标能够更好的符合网络的结构特点。同理，在电力网络中，平均最短路径为 16，此时只考虑三阶路径的 LP 指标就不够精确了。关于平均最短路径和考虑的最优路径长度的关系在文献[36]中有详细讨

论。

AUC	PPI	NS	Grid	PB	INT	USAir
LP	0.970	0.988	0.697	0.941	0.943	0.960
LP*	0.970	0.988	0.697	0.939	0.941	0.959
Katz	0.972	0.988	0.952	0.936	0.975	0.956
LHN-II	0.968	0.986	0.947	0.769	0.959	0.778
Precision	PPI	NS	Grid	PB	INT	USAir
LP	0.734	0.292	0.132	0.519	0.557	0.627
LP*	0.734	0.292	0.132	0.469	0.121	0.627
Katz	0.719	0.290	0.063	0.456	0.368	0.623
LHN-II	0	0.06	0.005	0	0	0.005

表 4 基于路径的相似性指标。

另外，在计算复杂度方面，由于 LP 指标只考虑局部信息，其计算复杂度比考虑全局信息的 Katz 和 LHN-II 要小很多。LP 的计算复杂度约为 $O(N\langle k\rangle^3)$，而 Katz 指标和 LHN-II 指标的计算复杂度均为 $O(N^3)$。可见对于规模巨大（N 大）且较稀疏（平均度 $\langle k\rangle$ 小）的网络 LP 指标在计算速度上具有明显的优势。

3.3 基于随机游走的相似性指标

一类相似性算法是基于随机游走定义的，包括平均通勤时间（Average Commute Time）[37]，Cos+指标[38]，有重启的随机游走（Random Walk with Restart）[39]，SimRank 指标[40]，以及新提出的两种基于局部随机游走的指标[36]。下面逐一介绍

（1）平均通勤时间（Average Commute Time）简称 ACT。设 $m(x,y)$ 为一个随机粒子从节点 x 到节点 y 平均需要走的步数，那么节点 x 和 y 的平均通勤时间定义为

$$n(x,y)=m(x,y)+m(y,x)$$

其数值解可通过求该网络拉普拉斯矩阵的伪逆 L^+ 获得[37]。即：

$$n(x,y)=M(l^+_{xx}+l^+_{yy}-2l^+_{xy})$$

其中 l^+_{xy} 表示矩阵 L^+ 中相应位置的元素。可以说如果两个节点的平均通勤时间越小，那么两个节点越接近。通常网络被观察到有普遍的集聚效应，因此相隔较近的节点更容易连边。由此定义基于 ACT 的相似性为 （在此可忽略常数 M）：

$$s^{ACT}_{xy}=\frac{1}{l^+_{xx}+l^+_{yy}-2l^+_{xy}}。$$

（2）基于随机游走的余弦相似性（Cos+）。在由向量 $v_x=\Lambda^{\frac{1}{2}}U^T\vec{e}_x$ 展开的欧式空间内，L^+

中的元素 L^+_{xy} 可表示为两向量 v_x 和 v_y 的内积，即 $L^+_{xy}=v_x^T v_y$，其中 U 是一个标准正交矩阵，是由 L^+ 特征向量按照对应的特征根从大到小排列所得，Λ 为以特征根为对角元素的对角矩阵，T 表示矩阵转置，$\vec{e}_x$ 表示一个一维向量且只有第 x 个元素为 1，其它都为 0。由此定义余弦相似性[38]：

$$s_{xy}^{\cos+}=\cos(x,y)^+=\frac{L^+_{xy}}{\sqrt{L^+_{xx}\cdot L^+_{yy}}}$$

（3）重启的随机游走（Random Walk with Restart）简称 RWR。这个指标可以看成是网页排序算法（PageRank）的拓展应用[39]。它假设随机游走粒子在每走一步的时候都以一定概率返回初始位置。设粒子返回概率为 1-c，P 为网络的马尔科夫概率转移矩阵，其元素 $P_{xy}=a_{xy}/k_x$ 表示节点 x 处的粒子下一步走到节点 y 的概率，其中如果 x 和 y 相连则 $a_{xy}=1$，否则为 0。某一粒子初始时刻在节点 x 处，那么 t+1 时刻该粒子到达网络各个节点的概率向量：

$$\vec{q}_x(t+1)=c\cdot P^T\vec{q}_x(t)+(1-c)\vec{e}_x$$

其中 $\vec{e}_x$ 表示初始状态（其定义与 Cos+ 中相同）。不难得到上式的稳态解为 $\vec{q}_x=(1-c)(I-cP^T)^{-1}\vec{e}_x$，其中元素 q_{xy} 为从节点 x 出发的粒子最终有多少概率走到节点 y。由此定义 RWR 相似性为

$$s_{xy}^{RWR}=q_{xy}+q_{yx}$$

关于 RWR 的一种快速算法参见文献[41]。该指标已被应用于推荐系统的算法研究中[42]。

（4）SimRank 指标，简称 SimR。它的基本假设是如果两节点所连接的节点相似那么这两个节点就相似[40]。它的自洽定义式为

$$s_{xy}^{SimR}=C\frac{\sum_{z\in\Gamma(x)}\sum_{z'\in\Gamma(y)}s_{zz'}^{SimR}}{k_xk_y}$$

其中假定 $s_{xx}=1$，$C\in[0,1]$ 为相似性传递时的衰减参数。SimR 指标可以用来描述两个分别从节点 x 和 y 出发的粒子多久会相遇。

（5）局部随机游走指标（Local Random Walk）简称 LRW [36]。该指标与上述四种基于随机游走的相似性不同，它只考虑有限步数的随机游走过程。一个粒子 t 时刻从节点 x 出发，定义 $\pi_{xy}(t)$ 为 t+1 时刻这个粒子正好走到节点 y 的概率，那么可得到系统演化方程

$$\vec{\pi}_x(t+1)=P^T\vec{\pi}_x(t),\quad t\geq 0$$

其中 $\vec{\pi}_x(0)$ 为一个 $N\times1$ 的向量，只有第 x 个元素为 1，其它为 0，即 $\vec{\pi}_x(0)=\vec{e}_x$。设定各个节点的初始资源分布为 q_x，那么基于 t 步随机游走的相似性为

$$s_{xy}^{LRW}(t) = q_x \cdot \pi_{xy}(t) + q_y \cdot \pi_{yx}(t)。$$

刘伟平和吕琳媛给出了一种与度分布一致的初始资源分布，即 $q_x = k_x / M$ [36]，并在此基础上进行了大量实验，实验结果见表 6。LRW 相似性由于只考虑了有限步数的随机游走，此算法的计算复杂度相比较基于全局随机游走的 ACT，RWR，Cos+以及 SimR 算法都要小很多，因此对于规模较大，较稀疏的网络非常适用。

（6）叠加的局部随机游走指标（Superposed Random Walk）简称 SRW [36]。在 LRW 的基础上将 t 步及其以前的结果加总便得到 SRW 的值，即

$$s_{xy}^{SRW}(t) = \sum_{l=1}^{t} s_{xy}^{LRW}(l) = q_x \sum_{l=1}^{t} \pi_{xy}(l) + q_y \sum_{l=1}^{t} \pi_{yx}(l)$$

这个指标的目的就是给邻近目标节点的点更多的机会与目标节点相连。

文献[36]中比较了上述两种基于局部随机游走和基于全局随机游走的 ACT 和 RWR 指标在五个不同领域的网络中的链路预测效果。这五个网络分别为美国航空网络（USAir），科学家合作网（NS），电力网络（Grid），蛋白质相互作用网络（PPI）和线虫神经网络（C.elegans）。其拓扑结构的统计特性展现于表 5。注意，与 3.1 节中数据不同，这里只考虑了最大连通集。〈k〉和〈d〉分别表示平均度和平均最短距离。

Networks	**N**	**M**	**〈k〉**	**〈d〉**	**C**	**r**	**H**
USAir	332	2126	12.807	2.46	0.749	-0.208	3.464
NS	379	941	4.823	4.93	0.798	-0.082	1.663
Grid	4941	6594	2.669	15.87	0.107	0.003	1.450
PPI	2375	11693	9.847	4.59	0.388	0.454	3.476
C.elegans	297	2148	14.456	2.46	0.308	-0.163	1.801

表 5 五个网络的统计特征

表 6 总结了四种基于随机游走的相似性的链路预测精度。每个网络中精确度最大的值被标黑。括号中的数字表示 LRW 和 SRW 指标所对应的最优行走步数。可见除了 NS 网络以外 LRW 和 SRW 指标无论 AUC 还是 Precision 都好于 ACT 和 RWR 指标。而在 NS 网络中虽然 RWR 表现稍好，但是其计算复杂度远远大于 LRW 和 SRW 指标。由于 ACT 和 RWR 的计算复杂度为 $O(N^3)$，而 LRW 和 SRW 为 $O(N\langle k\rangle^n)$，其中 n 为随机游走步数。由此可以推算对于 NS 网络来说计算 RWR 的时间复杂度要比 SRW 慢 1000 多倍，而 AUC 只提高了千分之一。

AUC	USAir	NS	Grid	PPI	C. elegans
ACT	0.901	0.934	0.895	0.900	0.747
RWR	0.977	**0.993**	0.760	0.978	0.889
LRW	0.972 (2)	0.989 (4)	0.953 (16)	0.974 (7)	0.899 (3)
SRW	**0.978** (3)	0.992 (3)	**0.963** (16)	**0.980** (8)	**0.906** (3)
Precision	USAir	NS	Grid	PPI	C. elegans
ACT	0.49	0.19	0.08	0.57	0.07

RWR	0.65	**0.55**	0.09	0.52	0.13
LRW	0.64 (3)	0.54 (2)	0.08 (2)	**0.86** (3)	**0.14** (3)
SRW	**0.67** (3)	0.54 (2)	**0.11** (3)	0.73 (9)	**0.14** (3)

表 6 四种基于随机游走的算法比较

3.4 权重在链路预测中的作用

含权网络的链路预测是一个较重要的方向，但到目前为止还没有系统的研究工作，对于如何更好的应用权重的信息以提高链路预测的精确度还没有明确的答案。文献[43]中，吕琳媛和周涛将三种局部算法 CN，AA 和 RA 拓展为含权形式。定义如下：

$$s_{xy}^{WCN} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} w(x,z)^{\alpha} + w(z,y)^{\alpha}$$

$$s_{xy}^{WAA} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{w(x,z)^{\alpha} + w(z,y)^{\alpha}}{\log(1+s(z))}$$

$$s_{xy}^{WCN} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{w(x,z)^{\alpha} + w(z,y)^{\alpha}}{s(z)}$$

其中，$w(x,y)$ 为连接节点 x 和 y 的连边的权重，$s(x)=\sum_{z\in\Gamma(x)} w(x,z)^{\alpha}$ 为节点 x 的强度，α 参数用来调节权重在预测中的作用大小。当 $\alpha=0$ 时，WCN，WAA 和 WRA 指标分别回到各自不含权的形式（参见 3.1 节中的定义）。在三个实际网络中应用这三种含权指标进行链路预测，结果发现在链路预测中也存在弱连接效应[44]，即给原来权重较低的边赋予较大的权重，而原来权重大的边赋予较小的权重（redistribution），用新的权重会得到更好的预测效果。图 1 展示了美国航空网络的预测结果。在美国航空网络中，城市机场代表节点，航线代表边，边的权重由两机场间航班的飞行频次决定。从图中得到三种含权算法的最优参数 α 均小于 0。这意味着原来权重大的边在链路预测中的作用变小了，而原来权重小的边作用反而增大了，即所谓的弱连接效应。

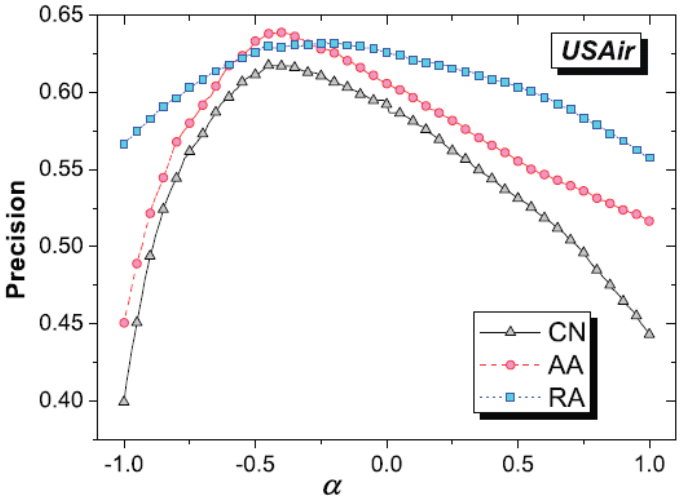

图 1 美国航空网络预测精度与参数 α 的关系。

此外在科学家合作网中也发现了弱连接效应。但是在 C.elegans 这个线虫的神经网络中结果恰恰相反，它的最优参数值均大于 1，这意味着只有更加强调强连接，弱化弱连接才

能得到更好的预测结果，可称之为链路预测的强连接效应。吕琳媛和周涛随后应用 motif 分析的方法从定性上解释了这种差异的原因，但是还不能进行定量化的描述。含权网络的预测方法研究还具有很大的拓展空间。同样的网络结构，不同的含权方式在实际预测中起到的作用也可能不一样。要搞清这些问题还需要更加深入细致的研究工作。

4. 基于最大似然估计的链路预测

链路预测另一类方法是基于最大似然估计的。Clauset, Moore 和 Newman[8]认为很多网络的连接可以看作某种内在的层次结构的反映，基于此，他们提出了一种最大似然估计的算法进行链路预测，这种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，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。Guimera 和 Sales-Pardo[10]假设我们观察到的网络是一个随机分块模型(Stochastic Block Model)[9]的一次实现，在该模型中节点被分为若干集合，两个节点间连接的概率只和相应的集合有关。他们所提出的基于随机分块模型的链路预测方法，可以得到比 Clauset, Moore 和 Newman 更好的结果。与此同时，该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。基于最大似然估计的方法一个最大的问题就是计算复杂度太高，因此并不适合在规模较大的网络中应用。

4.1 层次结构模型[8]。

对实际网络结构的实证研究表明，很多情况下网络具有一定的层次结构[30，45，46]。因此某个含有 N 个节点的网络可以由一个含有 N 个叶子节点和 N-1 个内部节点的树状图表示。每个内部节点赋予一个概率值 $p_r(\in[0,1])$，而两个结点相连接的概率就等于距离他们最近的共同祖先节点所赋予的概率。图 2 为一个用树形结构表示的含有 5 个节点的网络层次结构示意图。由此可见节点 1 和节点 2 连接的概率为 0.5，节点 1 和节点 3 连接的概率为 0.3，节点 3 与节点 4 连接的概率为 0.4。

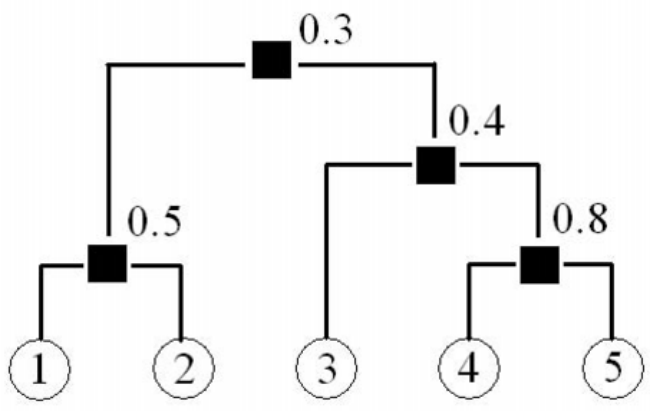

图 2 用树形图表示网络的层次结构示例。

给定一个网络 G，以及和它相对应的一个树形图 D。则这个树形图对目标网络 G 的似然估计值为

$$L(D,\{p_r\})=\prod_r p_r^{E_r}(1-p_r)^{L_r R_r-E_r}$$

其中 L_r 和 R_r 分别表示以内部节点 r 为根的左子树和右子树的叶子节点数目。 E_r 表示以 r 为最近共同祖先的节点对中有多少对在 G 中已形成了连边。对于给定的 D，使似然估计值最大的最优概率 $p_r^* = E_r/(L_r R_r)$，按此结果给 D 的每个内部节点赋概率值。对于网络 G 的多个树形图，$L(D,\{p_r\})$ 越大表示该树形图对网络的刻画越真切。由于能够得到最大似然估计值的树形图不只有一个，因此要考虑多个树形图的平均结果。采用马尔科夫链蒙特卡洛算法得到一组可用于链路预测的树形图。具体方法如下：

1） 首先给定一个树形图，并按照公式 $p_r^* = E_r/(L_r R_r)$ 给每个内部节点赋概率值。

2） 随机选择一个内部节点 r 并考虑两类子树，以其兄弟节点为根节点的子树集合 B，和以其儿女节点为根节点的子树集合 C。

3） 通过交换子树集合 B 和 C 中的子树获得新的树状图 D'，注意 D 和 D' 不同。

4） 从所有可能的 D' 中随机选择一个，当 $\log L(D') \geq \log L(D)$ 时接受新树状图 D'，否则以 $L(D')/L(D)$ 的概率接受 D'。然后从新回到第 2）步。

5） 当该马尔科夫链收敛于平稳的时候，开始生成可用的树形图，如 5000 个。

最终网络中未连边的两个节点 x 和 y 可能连边的概率为所有树形图中两节点连接概率的平均值 $\langle p_{xy} \rangle$。然后将所有未连边的节点对按照连接概率从大到小排列，排在最前面的表示出现连边的概率越大。

实验结果显示，此方法对于有明显层次结构的网络表现尚好，如恐怖袭击网络和草原食物链网络，而对于层次结构不明显的网络，如科学家合作网和线虫神经网络，表现还不如最简单的共同邻居算法好，具体比较参见文献[36]。另外，从链路预测实用性的角度来讲，这个方法的计算时间复杂度较大，通常使马尔科夫链收敛需要 $O(N^2)$ 步，而每一步都至少要执行上述步骤 2）至步骤 4）一次。因此不适用于规模较大的网络。

4.2 随机分块模型[10]。

随机分块模型也是一种基于最大似然估计的方法。它的基本思想是根据网络具有模块性的特点，将网络的节点分组，而每两个节点是否连边是由他们所在的组决定的。已知目标网络的节点数为 N，应用随机分块模型进行链路预测，首先需将 N 个节点分组，然后给每个组对赋予一个连接概率 $Q_{\alpha\beta}(\in[0,1])$，由此建立一个分块模型 M。根据此分块模型可以得到在组 α 内的节点 i 和在组 β 内的节点 j 连接的概率 $p(A_{ij}=1|M)=Q_{\alpha\beta}$。该分块模型对目标网络的可靠性为

$$p(A|M)=\prod_{\alpha\leq\beta} Q_{\alpha\beta}^{l_{\alpha\beta}}(1-Q_{\alpha\beta})^{r_{\alpha\beta}-l_{\alpha\beta}}$$

其中 A 为目标网络的邻接矩阵，$l_{\alpha\beta}$ 为原网络中组 α 内的节点与组 β 内的节点连边的数量，

$r_{\alpha\beta}$ 为组 α 内的节点与组 β 内的节点一共可连边的数量。可见此方法与层次结构模型的公式基本一致，其最优概率 $Q^*_{\alpha\beta} = l_{\alpha\beta} / r_{\alpha\beta}$。按上述方法生成所有可能的分块模型 M，最终由贝叶斯定理得到节点 i 和节点 j 的连边的可信度为

$$R^L_{ij} = p(A_{ij} = 1 | A) = \frac{\int_\Omega p(A_{ij} = 1 | M) p(A | M) p(M) dM}{\int_\Omega p(A | M') p(M') dM'}$$

其中 Ω 为所有可能的分块模型集合（实际运算中并不需要真正考虑所有的）。为方便计算可将 $p(M)$ 设定为一个常数。可信度越高表示越可能连边。随机分块模型不仅可以预测缺失边，还可以根据可信度判断哪些边是错误连边，例如对蛋白质相互作用关系的错误认识。随机分块模型平均而言表现比层次结构模型好，尤其在预测错误连边的时候。但是他与层次模型一样都存在计算时间复杂度高的问题。

5. 概率模型

应用概率模型进行链路预测的基本思路就是建立一个含有一组可调参数的模型，然后使用一些优化策略寻找最优的参数值使得所得到的模型能够更好的再现真实网络的结构和关系特征，然后网络中两个未连边的节点对连边的概率就等于在这组最优参数下它们之间产生连边的条件概率。如果将边的存在性（存在或不存在）看成是边的一种属性，那么链路预测问题就转变为预测边的属性问题[47]。两个常用的框架为概率关系模型框架（Probabilistic Relational Models）[48]和有向无环概率实体关系框架（Directed Acyclic Probabilistic Entity Relationship）[49]。他们的区别在于对数据库的表达方式不同，前者基于关系模型（Relational models），后者基于实体关系模型（Entity-relationship model）。

概率模型的优势在于较高的预测精确度，它不仅使用了网络的结构信息还涉及节点的属性信息。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。

5.1 概率关系模型（PRMs）

概率关系模型是将概率模型和关系模型相结合的一种预测模型。概率关系模型包括三个网络：1）数据网络（Data graph）即所谓的训练集，包含原始的数据信息；2）模型网络（Model graph），分析数据网络得到的用于刻画网络主体属性之间的关系，这种关系既包括一类主体内部属性之间的关系，也包括不同主体属性之间的关系。3）推理网络（Inference graph）是将模型网络与目标网络（测试集）相结合的网络，用于对目标网络的预测。根据模型网络的不同构造方法又可将概率关系模型分为：贝叶斯网络关系模型（Relational Bayesian Networks）[50]，马尔科夫网络关系模型（Relational Markov Networks）[51]和关系依赖网络模型（Relational Dependency Networks）[52,53]。

贝叶斯网络模型（RBN）[50]：贝叶斯网络 $G(V, E, P)$ 为一个有向无环图，由条件概率分布和网络结构两部分组成。其中 $V = \{v_1, v_3, \cdots, v_n\}$ 为节点集合，每个节点代表一个变量，即数据网络中所涉及的属性变量，E 为有向边集合，表示变量之间的关系，P 为一组条件概率，$p(v | pa(v))$ 表示节点 v 的父亲节 $pa(v)$ 点对他的影响。如果一条有向边从节

点 x 指向节点 y，那么节点 x 可视为节点 y 的父亲。于是该贝叶斯网络的联合概率分布为 $p(v_1, v_2, \cdots, v_n) = \prod_{i=1}^{n} p(v_i \mid pa(v_i))$。

马尔科夫网络模型（RMN）[51]：马尔科夫网络 $G(V, E, \Phi)$ 为一个无向图，允许环存在。其中 V 和 E 仍表示变量和变量关系集合，Φ 表示潜能函数。定义 C 为网络中完全子图的集合。每一个完全子图对应一个潜能函数 Φ_c，于是网络的联合概率分布为 $p(v_1, v_2, \cdots, v_n) = \frac{1}{Z}\prod_{c \in C} \Phi_c(v_c)$，其中 v_c 表示完全子图 c 中的节点，Z 为标准归一化函数。

关系依赖网络模型（RDN）[53]：关系依赖网络模型与上述两个模型的最大区别在于它不用优化整个联合概率分布，而是应用伪似然估计（Pseudo-likelihood [54]）的方法分别对每个变量的条件概率进行估计，也就是说在估计条件概率 $p(v \mid pa(v))$ 的时候并不考虑条件概率 $p(pa(v) \mid v)$。由于对变量的估计是独立的，相比 RBN 和 RMN 模型其计算复杂度降低很多。RDN 与 RBN 相似也是用有向图表示属性之间的依赖关系，但是他允许环的存在。

5.2 有向无环概率实体关系模型（DAPER）[49]

DAPER 是以实体关系模型为基础所建立的模型，它将实体之间的关系也看成是和实体一样重要。DAPER 模型包括六类组成成分：

（1） 实体类（Entity Classes），即网络的实体，如大学数据库中的学生类和课程类。

（2） 关系类（Relationship Classes），即描述实体间的关系，如学生选择课程中的选择关系。

（3） 属性类（Attribute Classes），即实体或者关系的属性，如学生的智商，课程的难易程度等。

（4） 弧线类（Arc Classes），用来描述各个属性之间的关系，如学生的课程分数受到学生智商和课程难度的影响。属性关系构成的网络为有向无环网络（与 RBN 类似）。

（5） 局部概率分布类（Local Distribution Classes），对某一属性类的条件概率分布，与 PRMs 中的条件概率类似。

（6） 限制条件类（Constraint Classes）衡量属性关系之间的限制条件。

Heckerman 在文[49]中比较了该模型与 PRMs 模型的区别和联系。图 3 展现了在学生选择课程的例子中分别用 DAPER（a）和 PRMs（b）建立的模型。

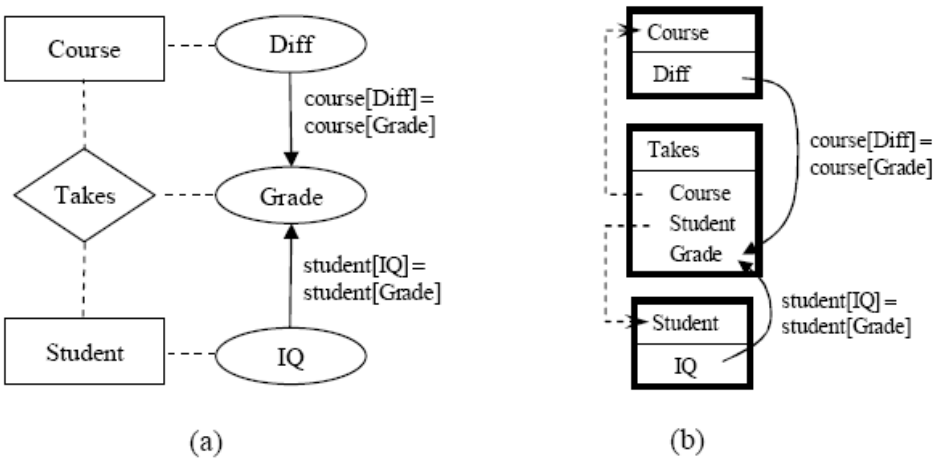

图 3 以大学数据中学生，课程为例的 DAPER 模型和 PRMs 模型。

6. 总结与展望

综上所述，无论是基于结构的相似性预测方法还是基于最大似然估计的方法，或是概率模型本质上都是通过对已知数据的尽可能真切的刻画来实现预测，但是他们的角度各自不同。基于结构相似性的方法只涉及网络的结构信息，它主要是从某一个角度对于网络的某一方面的结构特点进行刻画，如果目标网络的结构在这一方面特征显著，那么即可得到较好的预测效果。虽然基于网络结构相似性的方法比较简单，计算复杂度相对较低，特别是基于局部结构的算法，但是各个方法在不同网络中的预测能力大大不同。目前还没有算法性能和网络结构特征之间关系的较深入的研究。对于比较复杂的网络，例如含权网络、有向网络、多部分网络以及含有异质边的网络如何通过结构信息进行预测的讨论甚少且不系统[43，55，56]。基于最大似然估计的方法虽然也是基于网络结构的，但是它针对的是整个网络结构而不仅仅局限于某一方面。这类方法由于计算复杂度较高，不可能应用于规模较大的网络。实验显示这类方法的预测精度也不是很高。概率模型是数据挖掘的传统模型，他可以同时考虑网络的结构信息和节点的属性信息以求得到更好的预测效果。但是计算的复杂性以及节点外在属性信息在获取上的难度成为这类方法应用的局限性。

最近十年，复杂网络研究在很多科学分支，包括物理、生物、计算机等等掀起高潮[57]，其中相当一部分研究立足于揭示网络演化的内在驱动因素。仅以无标度网络(scale-free networks)为例[58]，已经报道的可以产生幂律度分布的机制就包括了富者愈富(rich-get-richer)机制[31]，好者变富(good-get-richer)机制[59]，优化设计(optimal design)驱动[60]，哈密顿动力学(Hamiltonian dynamics)驱动[61]，聚生(merging and regeneration)机制[62]，稳定性限制(stability constraints)驱动[63]，等等。可是，由于刻画网络结构特征的统计指标非常多，很难比较和判定什么样的机制能够更好再现真实网络的生长特性。利用链路预测有望建立简单的比较平台，能够在知道目标网络演化情况的基础上量化比较各种不同机制对于真实生长行为的预测能力，从而可以大大推动复杂网络演化机制的相关研究。

与此同时，受益于复杂网络研究的快速发展，基于网络结构的链路预测方法有望在网络理论的帮助下得到发展和完善。一方面如何以网络系综理论为基础，建立网络链路预测的理论框架，并产生对实际预测有指导作用的理论结论，例如通过对网络结构的统计分析估算各个方法的可预测的极限，从而指导选择最佳的预测方法等等；另一方面如何通过网络的结构信息借助复杂网络的分析工具设计高效的算法来处理大规模网络的链路预测问

题。

尽管已有一些论文讨论了如何将链路预测的方法和思想与一些应用问题，例如部分标号网络的节点类型预测[19, 64, 65]与信息推荐问题[33, 55, 66]，相联系的可能性与方法，但是，目前尚缺乏对于大规模真实数据在应用层面的深入分析和研究。这方面的研究不仅仅具有实用价值，而且有助于揭示链路预测这个问题本身存在的优势与局限性。

参考文献

[1]L. Getoor, C. P. Diehl, Link Mining: A Survey. ACM SIGKDD Explorations Newsletter, 2005, 7: 3.

[2]R. R. Sarukkai, Link prediction and path analysis using markov chains. Computer Networks, 2000, 33: 377.

[3]J. Zhu, J. Hong, J. G. Hughes, Using markov chains for link prediction in adaptive web sites. Lect. Notes Comput. Sci., 2002, 2311: 22.

[4]A. Popescul, L. Ungar, Statistical relational learning for link prediction. Proc. Workshop on Learning Statistical Models from Relational Data, New York: ACM Press, 2003: 81.

[5]J. O'Madadhain, J. Hutchins, P. Smyth, Prediction and ranking algorithms for even-based network data. Proc. ACM SIGKDD 2005, New York: ACM Press, 2005: 23.

[6]D. Lin, An information-theoretic definition of similarity. Proc. 15th Intl. Conf. Mach. Learn.. San Francisco, Morgan Kaufman Publishers, 1998: 296.

[7]D. Liben-Nowell, J. Kleinberg, The Link-Prediction Problem for Social Networks. J. Am. Soc. Inform. Sci. Technol., 2007, 58: 1019.

[8]A. Clauset, C. Moore, M. E. J. Newman, Hierarchical structure and the prediction of missing links in networks. Nature, 2008, 453: 98.

[9]P. W. Holland, K. B. Laskey, S. Leinhard, Stochastic blockmodels: First steps. Social Networks, 1983, 5: 109.

[10]R. Guimera, M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks. Proc. Natl. Sci. Acad. U.S.A., 2009, 106: 22073.

[11]H. Yu, et al., High-quality binary protein interaction map of the yeast interactome network. Science, 2008, 322: 104.

[12]M. P. H. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, C. Wiuf, Estimating the size of the human interactome. Proc. Natl. Sci. Acad. U.S.A., 2008, 105: 6959.

[13]L. A. N. Amaral, A truer measure of our ignorance. Proc. Natl. Sci. Acad. U.S.A., 2008, 105: 6795.

[14]L. Schafer, J. W. Graham, Missing data: Our view of the state of the art. Psychol. Methods, 2002, 7: 147.

[15]G. Kossinets, Effects of missing data in social networks. Social Networks, 2006, 28: 247.

[16]R. Kumar, J. Novak, A. Tomkins, Structure and evolution of online social networks. Proc. ACM SIGKDD 2006, New York: ACM Press, 2006: 611.

[17]B. Gallagher, H. Tong, T. Eliassi-Rad, C. Falousos, Using ghost edges for classification in sparsely labeled networks. Proc. ACM SIGKDD 2008, New York: ACM Press, 2008: 256.

[18]K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, A. Joshi, Social ties and their relevance to churn in mobile telecom networks. Proc. EDBT'08, New York: ACM Press, 2008: 668.

[19]C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Field, P. Bork, Comparative assessment of large-scale data sets of protein-protein interactions. Nature, 2002, 417: 399.

[20]R. Albert, A.-L. Barabasi, Statisrical mechanics of complex networks. Rev. Mod. Phys., 2002, 74: 47.

[21]S. N. Dorogovtsev, J. F. F.Mendes, Evolution of networks. Adv. Phys., 2002, 51: 1079.

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

[23]Y. Pan, D.-H. Li, J.-G. Liu, J.-Z. Liang, Detecting community structure in complex networks via node similarity. Physica A, 2010, 389: 2849.
[24]J. A. Hanely, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 1982, 143: 29.
[25]J. L. Herlocker, J. A. Konstann, K. Terveen and J. T. Riedl, Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst., 2004, 22: 5.
[26]T. Zhou, J. Ren, M. Medo, Y.-C. Zhang, Bipartite network projection and personal recommendation. Phys. Rev. E, 2007, 76: 046115.
[27]G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, Auckland: MuGraw-Hill, 1983.
[28]P. Jaccard, Etude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Science Naturelles, 1901, 37 : 547.
[29]T. Sorensen, 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. Biol. Skr., 1948, 5: 1.
[30]E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A.-L. Barabasi, Hierarchical Organization of Modularity in Metabolic Networks. Science, 2002, 297: 1553.
[31]A.-L. Barabasi, R. Albert, Emergence of scaling in random networks, Science, 1999, 286: 509.
[32]L. A. Adamic, E. Adar, Friends and neighbors on the web. Social Networks, 2003, 25: 211.
[33]T. Zhou, L. Lü, Y.-C. Zhang, Predicting missing links via local information. Eur. Phys. J. B, 2009, 71: 623.
[34]L. Lü, C.-H. Jin, T. Zhou, Similarity index based on local paths for link prediction of complex networks. Phys. Rev.E, 2009, 80: 046122.
[35]L. Katz, A new status index derived from sociometric analysis. Psychometrika, 1953, 18: 39.
[36]W.-P. Liu, L. Lü, Link Prediction Based on Local Random Walk. Europhys. Lett., 2010, 89: 58007.
[37]D. J. Klein, M. Randic, Resistance distance. J. Math. Chem., 1993, 12: 81.
[38]F. Fouss, A. Pirotte, J.-M. Renders, M. Saerens, Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. IEEE Trans. Knowl. Data. Eng., 2007, 19: 355.
[39] S. Brin, L. Page, The anatomy of a large-scale hypertextual Web search engine. Comput. Netw. & ISDN Syst., 1998, 30: 107.
[40]G. Jeh, J. Widom, SimRank: A measure of structural-context similarity. Proc. ACM SIGKDD 2002, New York: ACM Press, 2002: 271.
[41]H. Tong, C. Faloutsos, J.-Y. Pan, Fast random walk with restart and its applications, Proc. 6th Intl. Conf. Data Min., IEEE Press, 2006: 613.
[42]M.-S. Shang, L. Lü, W. Zeng, Y.-C. Zhang, T. Zhou, Relevance is more significant than correlation: Information filtering on sparse data. Europhys. Lett, 2009, 88: 68008.
[43]L. Lü, T. Zhou, Link prediction in weighted networks: The Role of Weak Ties. Europhys. Lett., 2010, 89: 18001.
[44]M. S. Granovetter, The strength of weak ties. Am. J. Sociology, 1973, 78: 1360.
[45]E. Ravasz, A.-L. Barabáasi, Hierarchical organization in complex networks. Phys. Rev. E, 2003, 67: 026112.
[46]C. Zhou, L. Zemanová, G. Zamora, C. C. Hilgetag, J. Kurths, Hierarchical organization unveiled by functional connectivity in complex brain networks. Phys. Rev. Lett., 2006, 97: 238103.
[47]B. Taskar, M.-F. Wong, P. Abbeel, D. Koller, Link prediction in relational data. Proc. of Neural Information Processing Systems (NIPS), Cambridge MA: MIT Press, 2003: 659.
[48]N. Friedman, L. Getoor, D. Koller, A. Pfeffer, Learning probabilistic relational models. Proc. 16th Intl. Joint Conf. Artif. Intell. (IJCAI), Stockholm, Sweden, 1999: 1300.
[49]D. Heckerman, C. Meek, D. Koller, Probabilistic entity-relationship models, PRMs, and plate models. Proc.

21st Intl. Conf. Mach. Learn., Banff, Canada, 2004: 55.
[50]D. Heckerman, D. Geiger, D. Chickering, Learning bayesian networks: the combination of knowledge and statistical data. Mach. Learn., 1995, 20: 197.
[51]B. Taskar, P. Abbeel, D. Koller, Discriminative probabilistic models for relational data. Proc. UAI2002, Edmonton, Canada, 2002: 485.
[52]D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, C. Kadie, Dependency networks for inference, collaborative filtering, and data visualization. J. Mach. Learn. Res., 2000, 1: 49.
[53]J. Neville, D. Jensen, Relational dependency networks. J. Mach. Learn. Res., 2007, 8: 653.
[54]J. Besag, Statistical analysis of non-lattice data. The Statistician, 1975, 24: 179.
[55]J. Leskovec, D. Huttenlocher, J. Kleinberg, Predicting positive and negative links in online social networks. Proc. WWW 2010, New York: ACM, 2010: 641.
[56]T. Murata, S. Moriyasu, Link prediction of social networks based on weighted proximity measures. Proc. IEEE/WIC/ACM Intl. Conf. Web Intelligence, New York: ACM Press, 2007: 85.
[57]A.-L. Barabasi, Scale-Free Networks: a decade and beyond. Science, 2009, 325: 412.
[58]G. Caldarelli, Scale-Free Networks: complex webs in nature and technology. New York: Oxford Press, 2007.
[59]D. Garlaschelli, A. Capocci, G. Caldarelli, Self-organized network evolution coupled to extremal dynamics. Nat. Phys., 2007, 3: 813.
[60]S. Valverde, R. F. Cancho, R. V. Sole, Scale-free networks from optimal design. Europhys. Lett., 2002, 60: 512.
[61]M. Baiesi, S. S. Manna, Scale-free networks from a Hamiltonian dynamics. Phys. Rev. E, 2003, 68: 047103.
[62]B. J. Kim, A. Trusina, P. Minnhagen, K. Sneppen, Self organized scale-free networks from merging and regeneration. Eur. Phys. J. B, 2005, 43: 369.
[63]J. I. Perotti, O. V. Billoni, F. A. Tamarit, D. R. Chialvo, S. A. Cannas, Emergent self-organized complex network topology out of stability constraints. Phys. Rev. Lett., 2009, 103: 108701.
[64]Q.-M. Zhang, M.-S. Shang, L. Lü, Similarity-based classification in partially labeled networks. Int. J. Mod. Phys. C, 2010, 21: 813.
[65]P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, T. Eliassi-Rad, Collective classification in network data. AI Magazine, 2008, 29: 93.
[66]T. Zhou, Statistical mechanics of information systems: information filtering on complex networks, Ph. D. Thesis, University of Fribourg, 2010.