复杂网络链路预测的研究现状及展望

吕琳媛

前言：做链路预测这个方向有一年多的时间了，有一些收获和体会。一直想写一个综述进行总结，总是希望这个综述尽可能的包括更多更全面的信息，但是新的思想和结果源源不断的涌现，所谓的综述也就无限期的搁置了下来。前不久刚刚和伟平合作发表了一篇关于利用网络局部随机游走进行链路预测的文章，借此文发表之动力，总结一下链路预测这个方向的研究进展以及展望。希望该文能对那些正奋战在这个方向上和希望在此领域有所建树的科研工作者有所帮助和启迪。

（本文中所提到的具体的技术方法以及实验结果将在另一篇中文综述中详细介绍。）

1．链路预测及其研究意义

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链接的可能性[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。该问题的研究在理论和应用两个方面都具有重要的意义和价值。

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。这也是我们相比计算机专业的人研究链路预测的优势所在。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[2, 3]。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[4]，这个问题和网络聚类等应用息息相关[5]。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。

链路预测研究不仅具有如上所述的理论价值，其更重要的意义还是体现在应用方面。很多生物网络，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[6]，而对于人类自身，我们知道的仅有可怜的 0.3%[7,8]。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[9,10]。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，即对未来

的预测。举例来说，近几年在线社交网络发展非常迅速[11]，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[12]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[13]。最近在一篇关于链路预测的工作中提到了不仅可以预测所谓的缺失链接还可以预测网络中的错误链接[14]，这对于网络重组和结构功能优化有重要的应用价值。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[15]，我们就有可能应用链路预测的方法对其进行纠正。

2．研究现状

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一些早期的研究。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[16]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[17]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[18]提出一个回归模型在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[19]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[20]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。因此与节点属性信息相比较，已观察到的网络结构或者用户的历史信息更容易获得也是更可靠的。

近几年，基于节点相似性的链路预测方法受到了广泛的关注。此方法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。因此如何定义节点的相似性就成为该方法的一个核心问题。尽管这个框架非常简单，但是相似性定义本身内涵丰富，它既可以是非常简单的共同邻居的个数，也可以是包含了复杂数学物理内容的诸如随机游走的平均通讯时间[21]或者是基于图论的矩阵森林方法 [22]。因此这个简单的框架事实上提供了无穷无尽的可能性。Liben-Nowell 和 Kleinberg[23]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。他们发现，在仅考虑节点邻居信息的若干指标中，Adamic-Adar 参数[24]表现最好。周涛、吕琳媛和张翼成[25]在 6 种不同网络中比较了 9 种已知的基于局部信息的相似性指标在链路预测中的效果，并提出了两种新指标：资源分配指标(resource allocation index)和局部路径指标(local path index)。研究发现，新提出来的这两种指标具有明显好于包括 Adamic-Adar 参数在内的 9 中已知指标的预测能力。最近其他小组的研究结果显示，新提出来的相似性指标在进行群落划分[5]和含权网络权重设置[26]的时候也比原有指标好。吕琳媛、金慈航和周涛[27]进一步在噪音强度以及网络密度可控的网络模型中细致分析了局部路径指标的性能，发现这个指标在网络的平均最短路径较小的时候具有与依赖于网络全局结构信息的指标，例如 Katz 参数[28]，可匹敌的预测能力，

甚至在噪声较大的情况下可以比 Katz 参数预测的更加准确。另外，由于局部路径指标仅仅考虑了网络的局部信息，其计算量远远小于基于全局信息的指标，特别是在网络规模较大且稀疏的情况下，局部路径指标在计算复杂度上的优势更加明显，因此其应用前景相当可观。最近，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，通过与其他五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，而最优的游走步数受到网络平均距离的强烈影响。此外，在五种网络上的比较结果显示该方法比 08 年 Nature[31]上提出的基于网络层次结构的预测方法准确度更高。另外，Huang 等人的实验结果显示[30]，在得到节点间的直接相似性后，利用协同过滤技术对相似性指标进行一轮加权处理，一般而言可以得到更好的结果。这一方法已广泛应用于推荐算法的设计上，并得到了成功。实际上，个性化推荐可以看作是链路预测的一个子问题。

链路预测另一类方法是基于最大似然估计的。Clauset, Moore 和 Newman[31]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[14]假设我们观察到的网络是一个随机分块模型(Stochastic Block Model)[32]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。以此同时，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。

另外一个需要特别注意的趋势，是随着一些原来从事复杂网络研究的学者对链路预测问题的关注，很多复杂网络，特别是社会网络分析中遇到的理论与方法被应用到链路预测中。例如吕琳媛和周涛[33]发现在针对某些含权网络进行链路预测的时候，权重很小的边反而起到了比高权重边更大的作用，这与社会网络研究中广为人知的“弱连接理论”[34]有深刻的关联。Leskovec, Huttenlocher 和 Kleinberg[35]则注意到了近期“社交平衡理论”的定量化研究成果[36,37]，并在此启发下设计了可以预测网络中的正负（友敌）链接的算法。

链路预测最近两年受到了比较多的关注，很可能得益于 Clauset, Moore 和 Newman 在 08 年发表的《自然》论文[31]，以及 Redner 在《自然》上的评论文章[38]。弗里堡小组较早地认识到链路预测问题的重要价值，并开展了一系列的工作。同时，通过大力的宣传国内对这个方向已经开始有一些关注。湘潭大学胡柯小组[40]利用链路预测方法预测人类蛋白质相互作用网络中的致病基因，也得到了不错的精度。最近胡柯小组及青岛理工大学许小可与弗里堡小组就有向网络的链路预测问题和社会平衡理论应用于链路预测的问题展开了紧密合作。

3．前沿趋势分析及展望

我们注意到一方面受阻于网络节点外在属性在获取上的难度，另一方面受益于复杂网络研究的快速发展，链路预测问题的主要研究热点逐渐从依赖于节点属性的方法转移到只利用网络结构信息的方法上[23]。显然，后者在理论上也更优美简洁。不过，这个方面的研究主要集中在社会网络上，对于大量算法在各种不同网络中的预测能力的系统分析的总结尚欠。另外，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、有向网络和多部分网络的讨论虽然有[33,35,41]，但非常少，也不系统。相关的研究应该是近几年该方向的主流。

网络系综理论以及与之关联的网络熵的概念以及最大似然估计方法有望推动形成复杂网络的统计力学理论基础[42-44]。这方面研究存在的一个问题是熵的精确计算复杂性非常

大[45]，对于大规模网络而言往往不能实现。最近的一些链路预测算法[14,31]已经应用了网络系综和最大似然的概念，但是这些算法计算复杂性很大，精确性也不是很高[29]，例如文献[31]的方法目前只能处理数千节点的网络，且其预测效果对于不具有明确层次结构的网络并不好[29]。我们认为以下两个问题应该是目前国际上相关研究小组比较关注的：一是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可预测的极限，指导选择不同的预测方法等等；二是如何设计高效的算法来处理大规模网络的链路预测问题。

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[46]，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[47]，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[48]，好者变富(good-get-richer)机制[49]，优化设计(optimal design)驱动[50]，哈密顿动力学(Hamiltonian dynamics)驱动[51]，聚生(merging and regeneration)机制[52]，稳定性限制(stability constraints)驱动[53]，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。Guimera 和 Sales-Pardo 在 09 年的 PNAS 中已经提到网络重建(network reconstruction)的问题，表达了相近的思想，但是这方面的研究尚未见详细的报道。

尽管有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[12,39,54]与信息推荐问题[25,35,55]，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。

综上所述可概括为以下五个方面：

1) 丰富和提高现有相似性预测的算法，特别是针对有向网络、含权网络、多部分网络、含异质边的网络等较复杂的情形，提出新的相似性指标；
2) 对已知算法的性能进行深入细致的分析，揭示算法性能和网络结构特征之间的关系，希望得到各种算法在不同网络中的可预测性极限；
3) 利用网络系综和最大似然估计的思想和技术，建立基于相似性框架的链路预测的理论基础；
4) 基于链路预测的思想，建立可以针对给定演化轨迹的目标网络后评价不同演化机制的平台；
5) 实现有代表性的链路预测的应用研究，并开展自适应性的快速算法研究以实现在巨大规模的实际系统中的应用。

致谢： 感谢周涛在本文撰写过程中提供的帮助。

参考文献

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

[2] R. Albert, A.-L. Barabasi, *Statisrical Mechanics of Complex Networks*, Rev. Mod. Phys. 74 (2002) 47.

[3] S. N. Dorogovtsev, J. F. F.Mendes, *Evolution of networks*, Adv. Phys. 51 (2002) 1079.
[4] E. A. Leicht, P. Holme, M. E. J. Newman, *Vertex similarity in networks*, Phys. Rev. E 73 (2006) 026120.
[5] Y. Pan, D.-H. Li, J.-G. Liu, J.-Z. Liang, *Detecting community structure in complex networks via node similarity*, Physica A (to be published).
[6] H. Yu, *et al.*, *High-quality binary protein interaction map of the yeast interactome network*, Science 322 (2008) 104.
[7] 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. 105 (2008) 6959.
[8] L. A. N. Amaral, *A truer measure of our ignorance*, Proc. Natl. Sci. Acad. U.S.A. 105 (2008) 6795.
[9] L. Schafer, J. W. Graham, *Missing data: Our view of the state of the art*, Psychol. Methods 7 (2002) 147.
[10] G. Kossinets, *Effects of missing data in social networks*, Social Networks 28 (2006) 247.
[11] R. Kumar, J. Novak, A. Tomkins, *Structure and evolution of online social networks*, Proc. ACM SIGKDD 2006, ACM Press, New York, 2006, p. 611.
[12] B. Gallagher, H. Tong, T. Eliassi-Rad, C. Falousos, *Using ghost edges for classification in sparselylabelednetworks*, Proc. ACM SIGKDD 2008, ACM Press, New York, 2008, p. 256.
[13] 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, ACM Press, New York, 2008, p. 668.
[14] R. Guimera, M. Sales-Pardo, *Missing and spurious interactions and the reconstruction of complex networks*, Proc. Natl. Sci. Acad. U.S.A. 106 (2009) 22073.
[15] 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 417 (2002) 399.
[16] R. R. Sarukkai, *Link prediction and path analysis using markov chains*, Computer Networks 33 (2000) 377.
[17] J. Zhu, J. Hong, J. G. Hughes, *Using markov chains for link prediction in adaptive web sites*, Lect. Notes Comput. Sci. 2311 (2002) 22.
[18] A. Popescul, L. Ungar, *Statistical relational learning for link prediction*, Proc. Workshop on Learning Statistical Models from Relational Data, ACM Press, New York, 2003, p. 81.
[19] J. O' Madadhain, J. Hutchins, P. Smyth, *Prediction and ranking algorithms for even-based network data*, Proc. ACM SIGKDD 2005, ACM Press, New York, 2005.
[20] D. Lin, *An information-theoretic definition of similarity*, Proc. 15th Intl. Conf. Machine Learning, Morgan Kaufman Publishers, San Francisco, 1998.
[21] 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. 19 (2007) 355.
[22] P. Chebotarev, E. Shamis, *The matrix-forest theorem and measuring relations in small social groups*, Automat. Remote Contr. 58 (1997) 1505.
[23] D. Liben-Nowell, J. Kleinberg, *The Link-Prediction Problem for Social Networks*, J. Am. Soc. Inform. Sci. Technol. 58 (2007) 1019.
[24] L. A. Adamic, E. Adar, *Friends and neighbors on the web*, Social Networks 25 (2003) 211.
[25] T. Zhou, L. Lü, Y.-C. Zhang, *Predicting missing links via local information*, Eur. Phys. J.

B 71 (2009) 623.
[26] Y.-L. Wang, T. Zhou, J.-J. Shi, J. Wang, D.-R. He, *Emipirical analysis of dependence between stations in Chinese railway network*, Physica A 388 (2009) 2949.
[27] L. Lü, C.-H. Jin, T. Zhou, *Similarity index based on local paths for link prediction of complex networks*, Phys. Rev.E 80 (2009) 046122.
[28] L. Katz, *A new status index derived from sociometric analysis*, Psychometrika 18 (1953) 39.
[29] W.-P. Liu, L. Lü, *Link Prediction Based on Local Random Walk*, Europhys. Lett.89 (2010) 58007.
[30] Z. Huang, X. Li, H. Chen, *Link prediction approach to collaborative filtering*, Proc. 5th ACM/IEEE-CS Joint Conf. Digital Libraries, ACM Press, New York, 2005.
[31] A. Clauset, C. Moore, M. E. J. Newman, *Hierarchical structure and the prediction of missing links in networks*, Nature 453 (2008) 98.
[32] P. W. Holland, K. B. Laskey, S. Leinhard, *Stochastic blockmodels: First steps*, Social Networks 5 (1983) 109.
[33] L. Lü, T. Zhou, *Link Prediction in Weighted Networks: The Role of Weak Ties*, Europhys. Lett. 89 (2010) 18001.
[34] M. S. Granovetter, *The strength of weak ties*, Am. J. Sociology 78 (1973) 1360.
[35] J. Leskovec, D. Huttenlocher, J. Kleinberg, *Predicting Positive and Negative Links in Online Social Networks*, Proc. WWW 2010, ACM, New York, 2010.
[36] T. Antal, P. Krapivsky, S. Redner, *Dynamics of social balance on networks*, Phys. Rev. E 72 (2005) 036121. [37] S. Marvel, S. Strogatz, J. Kleinberg, *Energy landscape of social balance*, Phys. Rev. Lett. 103 (2009) 198701.
[38] S. Redner, *Teasing out the missing links*, Nature 453 (2008) 47.
[39] Q.-M. Zhang, M.-S. Shang, L. Lü, *Similarity-based classification in partial labeled networks*, arXiv: 1003.0837.
[40] L. Zhang, K. Hu, Y. Tang, *Predicting disease-related genes by topological similarity in human protein-protein interaction network*, Cent. Eur. J. Phys. (to be published).
[41] T. Murata, S. Moriyasu, *Link prediction of social networks based on weighted proximity measures*, Proc. IEEE/WIC/ACM Intl. Conf. Web Intelligence, ACM Press, New York, 2007.
[42] G. Bianconi, *Entropy of network ensembles*, Phys. Rev. E 79 (2009) 036114.
[43] K. Anand, G. Bianconi, *Entropy measures for networks: Toward an information theory of complex topologies*, Phys. Rev. E 80 (2009) 045102.
[44] G. Bianconi, P. Pin, M. Marsili, *Assessing the relevance of node features for network structure*, Proc. Natl. Acad. Sci. U.S.A. 106 (2009) 11433.
[45] J. Li, B.-H. Wang, W.-X. Wang, T. Zhou, *Network Entropy Based on Topology Configuration and Its Computation to Random Networks*, Chin. Phys. Lett. 25 (2008) 4177.
[46] A.-L. Barabasi, *Scale-Free Networks: A Decade and Beyond*, Science 325 (2009) 412.
[47] G. Caldarelli, *Scale-Free Networks: Complex webs in nature and technology*, Oxford Press, New York, 2007.
[48] A.-L. Barabasi, R. Albert, *Emergence of scaling in random networks,* Science 286 (1999) 509.
[49] D. Garlaschelli, A. Capocci, G. Caldarelli, *Self-organized network evolution coupled to extremal dynamics,* Nat. Phys. 3 (2007) 813.
[50] S. Valverde, R. F. Cancho, R. V. Sole, *Scale-free networks from optimal design,* Europhys. Lett. 60 (2002) 512.
[51] M. Baiesi, S. S. Manna, *Scale-free networks from a Hamiltonian dynamics,* Phys. Rev. E 68 (2003)

047103.
[52] B. J. Kim, A. Trusina, P. Minnhagen, K. Sneppen, *Self organized scale-free networks from merging and regeneration*, Eur. Phys. J. B 43 (2005) 369.
[53] 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. 103 (2009) 108701.
[54] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, T. Eliassi-Rad, Collective classification in network data, AI Magazine 29 (2008) 93.
[55] T. Zhou, *Statistical Mechanics of Information Systems: Information Filtering on Complex Networks*, Ph. D. Thesis, University of Fribourg, 2010.