网络大数据
——复杂网络的新挑战：如何从海量数据获取信息？

周 涛

(电子科技大学互联网科学中心 成都 610054)

doi:10.3969/j.issn.1001-0548.2013.01.004

2012年3月，奥巴马政府公布了“大数据研发计划”，美国国家科学基金会、国防部、能源部、国家健康研究所、地质勘探局和国防部先进研究计划局六个联邦部门和机构共同投资2亿美元，致力于提高和改进人们从海量和复杂的数据中获取知识的能力。这是美国1993年宣布“信息高速公路”计划后又一次重大科技发展部署。2012年5月，我国召开第424次香山科学会议，这是我国第一个以大数据为主题的重大科学工作会议。中国计算机学会、通信学会等于今年分别成立了“大数据专家委员会”。国家自然科学基金委员会2013年的《项目指南》中，大数据成为最热门关键词！2012年12月13日，中关村成立大数据产业联盟，由云基地、联通、用友、联想、百度、腾讯、阿里巴巴等企业组成了第一批理事单位。

数据量的激增带来了很多共性问题，譬如数据的可表示、可处理和可靠性问题等等。与此同时，各学科自身也有各具特色的大数据问题。网络科学既是以网络为研究对象的一门有数百年历史的专业性很强的学科，又是众多学科中不同研究对象的统一抽象的表达方式，其所遭遇的问题和挑战往往特别典型、特别重要！目前万维网具有超过万亿的统一资源定位符(URL)，Facebook有10亿节点和千亿连边，大脑神经元网络有数百亿节点，中国三大运营商的手机通讯网络无一不拥有数亿用户……如何处理超大规模的网络数据，已经成为学术界和企业界亟待解决的关键科学技术问题。

很多与网络紧密相关的大数据问题是具有共性的。网络数据是典型的非结构化数据，针对大型网络的存储和管理的图数据库设计是目前非关系型数据库的一个重要分支。尽管有学者坚信随着计算能力和数据采集能力的提升，处理全体数据将成为趋势，但抽样仍然是目前处理海量数据问题的一种常用方法，而网络抽样不同于从一堆数中抽样去逼近原始分布，后者有明确的最优目标，前者则无章可循——什么样的网络抽样才算是好的呢？应该用什么方法抽样呢？抽样误差如何估计呢？大数据之间需要通过关联和交叉复用展现出1+1>2的价值，以网络科学的语言来做比喻，就是希望破译“人人网”里面的某A就是“中国移动手机通讯网络”中的某B，并且分析两个网络之间到底存在多少结构和功能的关联性。另外，可视化展示能够帮助科学家快速从大数据中验证科学猜想并获得新的科学发现，大规模网络的可视化也已被认为是一种有助于理解和分析网络的有效方法。

除了上述提到的一些共性问题外，此处我们着重介绍两个网络大数据独特的问题：一是预测问题，二是图的快速算法问题。

预测是大数据最核心的科学问题。目前学术界主要关心两类预测问题，一是趋势预测，二是缺失信息预测。趋势预测是指通过事物的一些基本属性信息和早期的态势分析，预测事物发展的轨迹和最终影响力[1-2]。这样的例子很多，譬如通过分析社交网络中注册一个月的用户的行为以及这些用户与其他用户的互动，预测哪些用户将来会成为很有影响力的用户；通过用户–商品二部分图中产品的早期表现，例如一首新歌或一个新歌手上线一周的情况，来预测这首歌或者这个歌手有没有可能走红；通过一条信息早期数小时在微博网络上的传播情况，来预测这条信息最终的影响力等等。信息传播的趋势预测是一个正问题，其相应的反问题是对传播路径进行还原，确定扩散源节点的位置[3]。这个问题虽

收稿日期：2012 – 12 – 15

作者简介：周涛(1983 –)，男，教授，主要从事统计物理与复杂性科学方面的研究.

然不属于典型的预测问题，但也是相关且值得关注的问题。缺失信息预测假设我们观察到的网络只是真实网络的一部分，在这个基础上探讨如何利用当前信息去预测缺失边[4]。以基因调控网络和蛋白质相互作用网络为例，我们已经知道的网络结构只是完整结构很小的一部分，这时候缺失预测方法就能够起到很大的作用。另外，社交网络朋友推荐也可以看做是缺失信息预测，因为我们推荐的基本假设是“他们应该认识并成为好朋友”，其方法论和缺失信息预测是完全一致的。推荐系统设计的核心问题，就是用户–商品二部分图上的缺失信息预测[5]。这和上面提到的一部分图上的链路预测问题理念相近但方法技术上有所不同。

图的快速算法问题在大数据时代尤其具有挑战性。以前$O(N^2)$或者$O(N^3)$的算法就被认为效率很高了，而在动辄数亿节点的网络中，$O(N\log N)$甚至线性算法可能都是不可接受的——快速算法和分布式计算是必然的努力方向。在这种规模的网络上，即便是求取簇系数和平均距离，都是一件开销昂贵的事情。当然，这些毕竟还是简单的事情，因为精确计算的复杂性也不大，而且近似算法设计也比较容易。此处主要介绍图匹配的问题和图社区划分问题，因为这两个问题本身复杂性高，而且具有特别重要的应用价值。图匹配最严格的是要求判定两个同阶图是否同构，较弱的定义是判定两个图是否是子图同构的，也就是是否存在顶点之间的一个单射关系，若图A中两个顶点相连，则其在图B中的单射的两个顶点也必须相连。注意，此时A、B两个图不需要同阶，A的顶点数可以少于B。一般而言，两个图既不是同构的，也不会是子图同构的，这个时候，可以通过寻找最大公共诱导子图来描述两个图的相似性。这些问题在大数据时代往往没有太大实用价值，因为计算复杂性大得惊人，这个时候寻找近似的最大公共子图或者通过传播算法以及谱算法快速寻找两个图的顶点对应关系就变得重要了[6]。社区挖掘的重要性不需赘述，不仅是展开网络中观结构从而观察网络组织规律的有力武器，也对包括推荐系统设计[5]在内的很多网络应用问题的重要辅助算法。目前，表现良好的算法已经可以在单机上实现数小时内划分千万节点规模的简单无向网络[7]，划分效果主要还是采用模块化程度这一指标，尽管这个指标在社区规模分辨率等方面存在缺陷。社区挖掘还有一个针对超大网络非常直接的应用，就是大规模网络的分布式存储。这个时候我们希望把网络的节点分别存在在不同机器上，并且跨机器的交叉边越少越好，而且为了负载平衡，还要求每个机器上节点总数是差不多的。这就相当于社区挖掘的时候给出了两个限定条件，一是知道社区数目，二是要求每个社区的节点数几乎相等。最近微软亚洲研究院提出了一个可以处理十亿规模的分布式算法[8]。一个大胆的猜测是，现在和将来优秀的快速社团挖掘算法，也包括求解平均距离和其他网络特征的近似算法，都会越来越多地利用重整化群的理念与方法。

参 考 文 献

[1] ASUR S, HUBERMAN B A. Predicting the future with social media[C]//IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). New York: IEEE Press, 2010: 492-499.

[2] ALTSHULER Y, PAN W, PENTLAND A. Trends prediction using social diffusion models[J]. Lect Notes Comput Sci, 2012(7227): 97-104.

[3] PINTO P C, THIRAN P, VETTERLI M. Locating the source of diffusion in large-scale networks[J]. Phys Rev Lett, 2012(109): 068702.

[4] LÜL, ZHOU T. Link prediction in complex networks: a survey[J]. Physica A, 2011(390): 1150-1170.

[5] LÜL, MEDO M, YEUNG C H, et al. Recommender systems[J]. Physics Reports, 2012(519): 1-49.

[6] TIAN Y, MCEACHIN R C, SANTOS C, et al. SAGA: a subgraph matching tool for biological graphs[J]. Bioinformatics, 2007(23): 232-239.

[7] BLONDEL V D, GUILLAUME J-L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. J Stat Mech, 2008(10): 10008.

[8] WANG L, XIAO Y, SHAO B, et al. How to partition a billion-node graph[R]. Beijing: MSRA, 2012.

编 辑 蒋 晓