

网络科学热点问题荟萃

根据2012年《复杂性科学专栏》编委会会议决定,从2013年第一期起,每年第一期将组织网络科学同仁畅谈有关热点问题,为此本期刊登了三篇相关文章。第一篇“网络科学热点问题荟萃”由多位学者共同撰写,分别介绍网络统计、网络重构、网络优化、以及网络大数据问题。另有两篇:一篇介绍网络动力学,一篇介绍网络的网络。第二篇“共演博弈下网络合作动力学研究进展”由荣智海等人撰写,详细介绍了网络进化博弈中的三个重要方面进展。第三篇“相互依存网络鲁棒性研究综述”由李国颖等人撰写,介绍了相互依存网络的概念和模型,并简单描述了相互依存网络上的级联失效过程。我们相信通过这些网络科学热点问题介绍将有助于大家开阔视野,促进我国网络科学发展。

网络统计——复杂网络基础问题:为标度律提供统计支持

陈庆华¹, 史定华²

(1. 福建师范大学 福州 350007; 2. 上海大学数学系 上海 宝山区 200444)

doi:10.3969/j.issn.1001-0548.2013.01.001

复杂网络科学^[1]和与之相关的人类动力学^[2]发现,许多复杂现象存在令人惊奇的普适规律,即涌现标度特性。例如,复杂网络的度分布呈现 $P(k) \sim k^{-\gamma}$ 的幂律形式,人类时空行为往往阵发而非随机,导致任务等待时间也呈现这一形式。物理学家侧重探索它们形成的动力学机制,如增长择优连线或任务优先权规则等。然而确认标度律是否成立还同时需要得到广泛的统计支持。文献[3]就用一幅图展示了标度律的两大支柱:机制成熟和统计支持的景观。

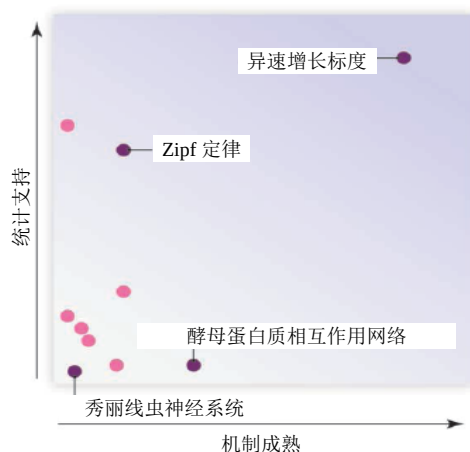


图1 统计支持

统计学是一门古老的学科,也是统计物理的基础,取得了许多重要成果,有着广泛的应用。在我国高校,统计学已列入一级学科,人才供不应求。科学出版社出版的《统计手册》^[4]有百余万字厚达千页,但却没有涉及标度律的统计方法。究其原因主要是定义不够明确,幂律形式是指近似幂律关系,严格幂律分布还是重尾分布,认识并不统一。

网络科学刚刚出现时,人们普遍采用简单的图估计方法进行统计。而且主要是在双对数坐标上画频率图,看上去基本成一条直线就认为服从标度律。由于大度数节点稀少,尾部摆动太大,就采用粗粒化的对数盒子图。在确定几何增长网络度分布指数时出现了反复,人们才开始使用画补分布图。关于这三种图的画法与比较,详细讨论参见文献[5]。画补分布图实质上等价于画秩次图,秩次图早在研究人类语言规律时就被采用。Zipf在研究文本中单词出现频次时,将所有单词排序,用横坐标表示序号,纵坐标表示对应的频次,在双对数坐标上画图得到了著名的Zipf标度律。文献[6]用两套数据比较了频率图和秩次图的优劣,数据1和数据2分别从幂律分布和指数分布产生,并分别画在双对数坐标和半对数坐标上,他们发现按秩次画图正确,按频率画图会出现错误。可见正确统计方法的重要性。

图估计方法全凭经验,结果因人而异。现在考虑数值方法,对于严格连续幂律分布:

$$p(x; \gamma) = \frac{\gamma - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\gamma} \quad x \geq x_{\min} > 0 \quad \gamma > 1 \quad (1)$$

如果得到一组数据 $(x_1, x_2, \dots, x_n)$, 将其看成随机样本, 文献[7]给出了 γ 的估计为:

$$\hat{\gamma} = 1 + n \left[\sum_{i=1}^n \ln(x_i / x_{\min}) \right]^{-1} \quad (2)$$

这是参数 γ 的极大似然估计, 它是有偏估计, 因为 $E[\hat{\gamma}] = 1 + [n(\gamma - 1)/(n - 1)]$ 。文献[8]证明了针对连续幂律分布不存在有效无偏估计, 只存在渐近有效无偏估计:

$$\tilde{\gamma} = 1 + (n - 1) \left[\sum_{i=1}^n \ln(x_i / x_{\min}) \right]^{-1} \quad (3)$$

进一步, 文献[8]还讨论了假设检验和似然比检验。引入连续伽玛分布 $G(x; r, \lambda)$:

$$p(x) = [\lambda^r / \Gamma(r)] x^{r-1} e^{-\lambda x} \quad x > 0, r > 0, \lambda > 0 \quad (4)$$

充分统计量 $T = \sum_{i=1}^n \ln(x_i / x_{\min})$ 服从 $G(x; n, \gamma - 1)$ 。对

于原假设 $H_0: \gamma = \gamma_0$ 对备择假设 $H_1: \gamma \neq \gamma_0$, 存在水平 α 的一致最优势检验, 拒绝域 $W = \{c_1 \leq T \leq c_2\}$ 。

由下面方程组确定:

$$\int_{c_1}^{c_2} G(x; n, \gamma_0 - 1) dx = \alpha$$

$$\int_{c_1}^{c_2} \{x - [n/(\gamma_0 - 1)]\} G(x; n, \gamma_0 - 1) dx = 0 \quad (5)$$

对于原假设 $H_0: \gamma = \gamma_0$ 对备择假设 $H_1: \gamma = \gamma_1$, 存在

水平为 α 的最优势检验, 两者的似然比统计量为

$$\lambda(x) = \prod_{i=1}^n p(x_i; \gamma_1) / \prod_{i=1}^n p(x_i; \gamma_0), \quad \text{拒绝域是 } W =$$

$\{T > c\}$, 其中 $c = G_{1-\alpha}(x; n, \gamma_0 - 1)$ 。虽然需要编程计算, 但这些数值结果比较客观公正。

然而, 复杂网络中的度是离散的, 上述估计结果需修正。另外, 实际网络的最小度也是随机变量, 而且对估计结果有重要影响。前面的讨论都是假定实际网络的度分布服从幂律分布, 而这是需要统计

检验的。更为重要的是度分布往往不是严格幂律的，而是重尾分布。重尾分布是一个很大的分布类，严格幂律分布只是其中最简单的代表。文献[9]发现Waring分布为：

$$P(k) = \frac{\beta(\beta+1)\cdots(\beta+k-1)\alpha}{(\beta+\alpha)(\beta+\alpha+1)\cdots(\beta+\alpha+k-1)(\beta+\alpha+k)} \quad \alpha, \beta > 0 \quad (6)$$

是另一个重要代表，几乎所有无标度增长网络模型的度分布都是Waring分布。如BA模型度分布是 $\alpha=2$ ， $\beta=m$ ；复制模型入度分布是 $\alpha=1$ ， $\beta=1$ 的Waring分布。因此，深入研究双参数Waring分布的统计分析方法具有重要的理论意义和实际价值。

参 考 文 献

- [1] BARABÁSI A-L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999(286): 509-512.
- [2] BARABÁSI A-L. The origin of bursts and heavy tails in human dynamics[J]. Nature, 2005(435): 207-211.
- [3] STUMPF M P H, PORTER M A. Critical truths about power laws[J]. Science, 2012(335): 665-666.
- [4] 茆诗松, 王静龙, 史定华, 等. 统计手册[M]. 北京: 科学出版社, 2003.
MAO Shi-song, WANG Jing-long, SHI Ding-hua, et al. Statistical handbook[M]. Beijing: Science Press, 2003.
- [5] 史定华. 网络度分布理论[M]. 北京: 高等教育出版社, 2011.
SHI Ding-hua. Theory of network degree distributions[M]. Beijing: Higher Education Press, 2011.
- [6] LI L, ALDERSON D, DOYLE J C, et al. Towards a theory of scale-free graphs: definitions, properties, and implications [J]. Internet Math, 2005(2): 431-523.
- [7] CLAUSET A, ROHILIA S, NEWMAN M E J. Power-law distributions in empirical data[J]. SIAM Review, 2009(51): 661-703.
- [8] 陈庆华, 陈月萍, 史定华. 幂律分布的统计性质及其应用 [C]//第八届全国复杂网络学术会议. 南京: [出版者不详], 2012.
CHEN Qing-hua, CHEN Yue-ping, SHI Ding-hua. The statistic property of power-law distributions and its applications[C]//The 8th Chinese Conference on Complex Networks. Nanjing: [s.n.], 2012.
- [9] 史定华. 关于无标度网络[C]//第八届全国复杂网络学术会议. 南京: [出版者不详], 2012.
SHI Ding-hua. About scale-free networks[C]//The 8th Chinese Conference on Complex Networks. Nanjing: [s.n.], 2012.

编 辑 蒋 晓

网络重构

——复杂网络的反问题：从时间序列重构网络拓扑和权重

王文旭

(北京师范大学管理学院 北京 海淀区 100875)

doi:10.3969/j.issn.1001-0548.2013.01.002

探测网络结构是研究、控制网络动力学行为的基础。然而，目前仍有大量未知网络结构的系统需要探索。造成这一困难的原因主要有三点：

1) 很多网络节点之间不存在实体连接，而是通过某种无法直接测量的信号相互作用，典型的例子有各种社会关系网络，通常只能通过间接的办法获得网络结构。间接推断各种社会网络有重要的理论和现实意义，比如通过破坏恐怖分子关系网络从而达到反恐的目的。

2) 很多实际网络系统在不断演化，从而导致网络结构不断变化。典型的网络有万维网(WWW)、通讯网络和朋友关系网络，如Facebook, Google group等。采用网络蜘蛛探测费时费力，并且存在滞后效应，导致得到的网络结构失去了时效性。

3) 目前的探测技术无法达到直接检测节点之间的实体链接，例如神经网络。这种情况下主要用粗粒化方法将大量的神经元作为一个整体，观察神经元集团之间的连接情况。另一种是建立神经网络。这类网络是功能层面的表示，而不是底层的物理连接网络。

基于以上原因，如何从个体单元行为重构复杂网络结构成为了复杂网络研究的前沿，受到越来越多的关注。生物信息学中一个很重要的问题是如何从实验数据重构各种网络，例如如何从基因表达数据重构基因调控网络等^[1]。目前的重构方法有利用逆向工程(reverse engineering)方法重构非线性动力系统^[2]，基于节点对外界微扰的响应(response to perturbation)推断网络结构的方法^[3]，基于同步现象

构造能够收敛到原系统的镜像系统，从而推断网络结构的方法^[4]，以及基于噪声的相关性方法能够比较准确的重构网络结构^[5-6]。

被美国科技评论评为当年十大科技进展之一的压缩传感(compressive sensing)理论是由文献[7]提出的新的信息获取理论。该理论证明对可压缩信号可以通过远低于Nyquist采样频率的方式进行数据采样，仍然能够精确地恢复出原始信号。该理论目前在图像处理、医学成像、模式识别、无线通讯、天文学等很多领域受到高度关注。

文献[8-10]考虑到已有的网络重构方法存在的缺陷，提出了基于压缩映象算法的复杂网络重构方法，通过相对于网络规模很少量的数据能够准确地重构出多种类型的复杂网络结构，包括无向网络、权重网络、有向网络等。主要工作有：提出从时间序列重构非线性系统并且预测catastrophes的理论方法^[8]；提出重构非线性和混沌振子网络的普适方法^[9]；提出基于策略和收益时间序列重构博弈网络的方法^[10]。文献[11]提出了从时间序列寻找隐藏节点和确定隐藏节点邻居的方法；文献[12]将反问题与正问题结合，提出了一种只基于数据的预测时间序列同步可能性的方法。下面简单描述一下利用压缩映象算法重构网络拓扑的思路。

应用压缩传感理论通过少量测量值 y 恢复稀疏向量 x 的示意图如图1所示。一个惊人的结果是在 x 稀疏并且 Φ 满足约束条件的情况下， M 可以远小于 N ，并且 x 中非零元素的个数也小于 M 。信号重构过程在此条件下转化为一个优化问题，求解方法有最

收稿日期：2012-12-15

基金项目：国家自然科学基金青年项目(11105011)；青年千人计划

作者简介：王文旭(1979-)，男，教授，主要从事复杂系统与复杂网络方面的研究。

小L1范数法和迭代阈值法等。可以把压缩映射方法看成一个黑箱,只要能够把网络重构问题构造造成压缩映射可以处理的形式,就可以通过已有的优化方法基于少量数据通过时间序列重构复杂网络结构和权重。对于一种特定的动力学,需要寻找动力学和结构的内在联系,将 $\mathbf{x}$ 构造成节点的邻居向量,并且能够从时间序列得到 $\mathbf{y}$ 向量和 Φ 矩阵。例如对于博弈网络, $\mathbf{y}$ 可以通过个体的收益时间序列得到, Φ 矩阵可以通过个体策略的时间序列获得, $\mathbf{x}$ 就是个体的邻居。因为复杂网络天然稀疏的,因此保证了 $\mathbf{x}$ 向量的稀疏性,从而可以应用压缩映射方法通过少量的时间序列准确重构网络结构和权重。具体重构方法参见文献[8-12]。

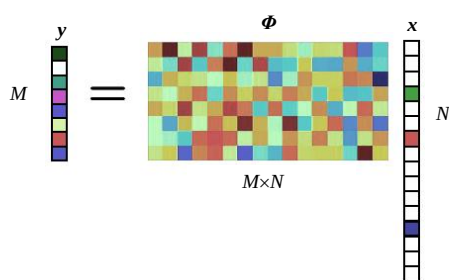


图1 基于压缩传感技术恢复信号示意图

图中, $\mathbf{x}$ 是长为 N 的待恢复稀疏向量, $\mathbf{y}$ 是长为 M 的测量信号, Φ 为 $M \times N$ 维满秩矩阵,满足约束等距性(restricted isometry property)条件。

目前,复杂网络的反问题,即从可观测量重构网络结构的工作还处于起步阶段,还有很多有待解决的问题。反问题与正问题不同,没有统一的研究范式,需要具体问题具体分析,针对不同的动力系统提出有针对性的重构方法。另外,将反问题与正问题结合是一个在未知网络结构前提下,研究复杂系统动力学行为的有效途径,将会有广阔的应用前景。

参考文献

- [1] GARDNER T S, d BERNARDO D, LORENZ D, et al. Inferring genetic networks and identifying compound mode of action via expression profiling[J]. Science, 2003, 301(5629): 102-105.
- [2] BONGARD J, LIPSON H. Automated reverse engineering of nonlinear dynamical systems[J]. Proc Natl Acad Sci USA, 2007, 104(24): 9943-9948.
- [3] TIMME M. Revealing network connectivity from response dynamics[J]. Phys Rev Lett, 2007, 98(22): 224101-1-4.
- [4] YU D, RIGHERO M, KOCAREV L. Estimating topology of networks[J]. Phys Rev Lett, 2006, 97(18): 188701-1-4.
- [5] WANG W-X, CHEN Q, HUANG L, et al. Scaling of noisy fluctuations in complex networks and applications to network prediction[J]. Phys Rev E, 2009, 80(1), 016116-1-6.
- [6] REN J, WANG W-X, LI B, et al. Noise Bridges Dynamical correlation and topology in coupled oscillator networks[J]. Phys Rev Lett, 2010, 104(5): 058701-1-4.
- [7] CANDÈS E J, ROMBERG J, Tao T. Stable signal recovery from incomplete and inaccurate measurements[J]. Comm Pure Appl Math, 2006, 59(8): 1207-1223.
- [8] WANG W-X, YANG R, LAI Y-C, et al. Predicting catastrophes in nonlinear dynamical systems by compressive sensing[J]. Phys Rev Lett, 2011, 106(15): 154101-1-4.
- [9] WANG W-X, YANG R, LAI Y-C, et al. Time-series based prediction of complex oscillator networks via compressive sensing[J]. EPL, 2011, 94(4): 48006-1-6.
- [10] WANG W-X, LAI Y-C, GREBOGI C, et al. Network reconstruction based on evolutionary-game data via compressive sensing[J]. Phys Rev X, 2011, 1(2): 021021-1-7.
- [11] SU R-Q, WANG W-X, LAI Y-C. Detecting hidden nodes in complex networks from time series[J]. Phys Rev E, 2012, 85(6): 065201-1-4.
- [12] SU R-Q, NI X, WANG W-X, et al. Forecasting synchronizability of complex networks from data[J]. Phys Rev E, 2012, 85(5): 056220-1-11.

编辑 蒋晓

网络优化——复杂网络设计问题：寻找最佳的网络结构

史定华

(上海大学数学系 上海 宝山区 200444)

doi:10.3969/j.issn.1001-0548.2013.01.003

网络科学从现实的复杂网络实证研究中发现了许多令人惊奇的规律,如小世界现象和无标度特性。于是,首先催生了网络建模热潮,再经过深入的网络分析和比较,最后都要回归于网络优化设计,以实现从认识世界到改造世界。最近Springer出版社推出了一本《复杂网络中的最优化手册》^[1],就是提醒网络科学研究者关注这个问题。

网络优化问题与网络实证研究相反,不是从收集实际网络数据出发,去统计度序列得到度分布;而是给定 N 个节点的度序列,去寻找具有某种功能的最优网络结构。它将面临三个子问题:讨论可实现性;研究实现方法;寻找最优结构。

1 给定度序列的图实现

现在任意给定一个度序列 $\{d_1, d_2, \dots, d_N\}$,其中 d_i 都是自然数。假定已按非增顺序排列,该度序列能否用简单无向图实现?注意:任给一个由自然数组成的有限序列,将它看作度序列并不能保证对应一个网络,图论中有如下重要定理。

Havel-Hakimi定理^[2] 度序列 $\{d_1, d_2, \dots, d_N\}$ 可用简单无向图实现当且仅当经过简化的度序列 $\{d_2-1, d_3-1, \dots, d_{d_1+1}-1, d_{d_1+2}, \dots, d_N\}$ 可用简单无向图实现。

递推应用(需重新排序)这个定理就可检测一个度序列是否可图实现。

如果度序列中给定度 k 的节点数 $n(k) \propto k^{-\gamma}$ (这里记号 $\propto$ 指正比),则该度序列称为幂律度序列^[3]。为了调查度指数 γ 与可图实现的相依关系,文献[4]从具有指数 $-2 \leq \gamma \leq 4$ 的有限幂律分布 $P(k) = k^{-\gamma} / H_{N-1, \gamma}$, $k = 1, 2, \dots, N-1$, 其中

$$H_{a,b} = \sum_{k=1}^a k^{-b}, \text{ 去随机生成范围在 } 1 \sim N-1 \text{ 的整数作}$$

为节点有幂律度序列的总体。通过数值模拟和极值分析得到:当幂律度序列指数 $0 \leq \gamma \leq 2$ 时一般不可图实现,而且指数0和2是两个一阶相变临界点。由此得出结论:“无标度网络(严格说指幂律图)都是稀疏的”。

2 度序列实现的多样性

除了按照Havel-Hakimi定理实现一个网络,还有许多其它实现方式。显然,给定齐性度序列 $d_1 = d_2 = \dots = d_N = d$,只要 Nd 是偶数,就有多种实现方式。例如,考虑度为3,某些节点数可图实现的网络,排除同构的多余网络,代数图论^[5]得到的不同网络数目如下。

节点数	网络数	节点数	网络数
4	1	14	509
6	2	16	4 060
8	5	18	41 301
10	19	20	510 489
12	85	26	2 094 480 864

可见随着网络规模增大,具有相同度序列的可能网络数目指数增长。

为了得到不同的网络结构,可采用简单的保度重连方法,即随机选择两条没有共用节点的连线 v_1-v_2, v_3-v_4 ;重新连线得到 v_1-v_3, v_2-v_4 或 v_1-v_4, v_2-v_3 。这种保持度序列不变的重连方法称为1阶零模型;对要求保持联合度分布的重连方法称为2阶零模型;进一步加以限制还有3阶零模型等等,详细讨论参见文献[6]。

文献[7]针对可图实现的度序列 $\{d_1, d_2, \dots, d_N\}$ 引

入了一个测度 $s(g) = \sum_{(i,j) \in E} d_i d_j$, g 表示实现度序列的

图, E 是图 g 的连线集合。显然, 对不同实现的图 g , 该测度的数值不同, 其中达到最大值 $s_{\max}$ 的那个图, 记为 $\tilde{g}$ 。这里就涉及网络最优化问题。

3 网络 L 矩阵的最优化

一个简单无向网络也可以用取值 0 (反映两个节点无连线) 和 1 (反映两个节点有连线) 的邻接矩阵 $A = [a_{ij}]$ 表示, 其中 $a_{ij} = a_{ji}$, $i, j = 1, 2, \dots$ 。令

$d_i = \sum_{j=1}^N a_{ij}$, $i = 1, 2, \dots, N$, 它就是度序列, 记

$D = \text{diag}\{d_1, d_2, \dots, d_N\}$, 矩阵 $L = D - A$ 称为拉普拉斯矩阵。由图谱理论有:

1) 简单无圈无权无向网络的 L 矩阵的特征值:

$$0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N;$$

2) 网络连通性的强弱与最小非零特征值有关, 即 $\lambda_2 > 0$ 则网络连通;

3) 网络生成树数目是 L 矩阵特征值之积除 N , 即

$$N_T = \prod_{i=2}^N \lambda_i / N;$$

4) 有界动力学网络的同步能力由特征比确定, 即 λ_2 / λ_N (无界为 λ_2)。

那么保持节点数和连线数不变, 什么网络结构同步能力最强? 或者更特殊些, 保持节点度序列不变, 什么网络结构同步能力最强? 通常, 对度齐性的网络, 特征比大时谱隙 λ_2 也大, 所以根据矩阵理论只需考虑下面最优化问题

$$\max_{A \in A^*} \frac{\lambda_2}{\lambda_N} = \max_{A \in A^*} \left\{ \min_{x^T e = 0, x \neq 0} \frac{x^T [D - A] x}{x^T x} \middle/ \max_{x^T e = 0, x \neq 0} \frac{x^T [D - A] x}{x^T x} \right\}$$

s.t. $d_1 = d_2 = \dots = d_N = d$, Nd 是偶数, $\lambda_2 > 0$, (1) 式中, Nd 是偶数保证可图实现; $\lambda_2 > 0$ 保证网络连通。

这个最优化问题是一个 NP-难题。文献[8]通过数值实验给出了一个解答^[8]。类似地有, 什么网络结构最容易控制? 什么网络结构最容易涌现合作行为? 什么网络结构抗攻击能力最强? 等等, 可见网络优化将会是网络科学的一个热点问题。

参考文献

- [1] THAI M T, PARDALOS P M (Editors). Handbook of optimization in complex networks: theory and applications [M]. New York: Springer, 2012.
- [2] HAKIMI S L. On the realizability of integers as the degrees of the vertices of a linear graph[J]. J Soc Ind Appl Math, 1962(10): 496-506.
- [3] CHUNG F, AIELLO W, LU L. A random graph model for power law graphs[J]. Experimental Math, 2001(10): 53-66.
- [4] GENIO C I, GROSS T, BASSLER K E. All scale-free networks are sparse[J]. Physical Review Letters, 2011(107): 178701.
- [5] MERINGER M. Fast generation of regular graphs and construction of cages[J]. J of Graph Theory, 1999(30): 137-146.
- [6] 汪小帆, 李翔, 陈关荣. 网络科学导论[M]. 北京: 高等教育出版社, 2012.
WANG Xiao-fan, LI Xiang, CHEN Guan-rong. Network science: an introduction[M]. Beijing: Higher Education Press, 2012.
- [7] LI L, ALDERSON D, DOYLE J C, et al. Towards a theory of scale-free graphs: definitions, properties, and implications[J]. Internet Math, 2005(2): 431-523.
- [8] SHI D H, CHEN G R, THONG W K, et al. Optimal homogeneous networks with best possible synchronizability [C]//Circuits and Systems(LASCAS). [S.l.]: [s.n.], 2012.

编辑 蒋晓

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许多共性问题，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不同机器上,并且跨机器的交叉边越少越好,而且为了负载均衡,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