幂律分布的文献小结

计算士
2010/6/3

看了一些 paper 后，一直想写一个关于幂律分布的文献综述，但近年来研究复杂系统，特别是复杂网络的文献增长迅速，而只要是涉及复杂网络的，基本上都谈到了类幂律分布，因此，这个计划因为耗时费力，就一直被搁置下来。

一直搁置也不是办法，就想先做一个 mini 版的，有时间再慢慢扩充，就会越来越全面。因此，就有了这个求精不求全的小版本。本文件收录的主题是幂律分布的**数学解析模型**和幂律分布**在实证数据中的拟合**，重点在前者，后者仅引用而不展开介绍。收录的都是我看过，觉得确实不错的论文，因为眼界有限，难免偏狭，有待继续完善。

Derek J. de Solla Price. Network of Science Papers. *Science* 149, 510-515 (1965).

这篇论文最早发现了科学论文引文网络中的幂律度分布，并提出了一个平衡态的链接模型，依靠论文平均引文数保持不变的假说，作者推出了一个引文增长网络结构，在该种结构下，被引用率高的文献的被引用次数会越来越多，因此最后形成论文被引用次数的幂律度分布。实际上，在这篇论文中，Price并没有给出详细的数学推导过程，有兴趣的同学可以参考K. K. Tung的*Topics in Mathematical Modeiling* (Princeton University Press, Cambridge, MA, 2007) 一书的第二章。

延伸：
科学论文引用网络的幂律度分布，数十年后，得到了更大规模数据的确认。可参考S. Redner. How popular is your paper*? An E*mpirical Study of the Citation Distribution. *The European Physical Journal B* 4, 131-134 (1998) 以及 Filippo Radicchi, Santo Fortunato, Benjamin Markines & Alessandro Vespignani. Diffusion of scientific credits and the ranking of scientists. *Physical Review E* 80, 056103 -056113 (2009).
此外，科学论文的合作者网络，也被证明是符合幂律度分布的，可参考 M. E. J. Newman. The structure of scientific collaboration networks. *PANS* 98, 404-409(2001) 以及 M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *PNAS* 101, 5200-5205(2004).

B. Mandelbrot. A Note On a Class of Skew Distribution Functions: Analysis and Critique of a Paper by H. A. Simon. *Information and control* 2, 90-99 (1959).

1959年到1960年，B. Mandelbrot和H. A. Simon在*Information and control*上有一场激烈的争论。两人都提出了自己的数学模型来解释语言学中的Zipf律，并指责对方的模型存在问题（Mandelbrot 1959，Simon 1960，Mandelbrot 1961a，Simon 1961a，Mandelbrot 1961b，Simon 1961b）。Simon认为Mandelbrot的“信息熵”的概念不适用于理解语言学中的信息概念，Mandelbrot指出自己的模型中的信息熵可以在非平衡态热力学和统计学的框架下得到更好的理解，代表的是一种最可能的状态，而并不局限于语言传递的真实信息，同时，Mandelbrot在这篇文章中指出Simon的Zipf模型只在幂指数小于-2的情况下成立，而且Simon模型中的“每次增加一个”的假设对于语言学是适用的，但并不能很好地解释经济学中的财富分布等情况。

Albert-Laszlo Barabasi & Reka Albert. Emergence of scaling in random networks. *Science* 286, 509-512(1999).

这篇论文提出了“优先链接”的模型，在这个模型中，每次新加入网络的节点倾向于把自己链接分配给已有较多链接的节点，这样，一个随机链接的初始网络最终将演化出具有幂律度分布结构的网络。作者提出了“无标度”（scale-free）的思想来理解幂律分布，并指出许多实际的网络，如互联网和电力网，都是无标度的。

延伸：
互联网链接的幂律度分布结构的实证研究，可参考 A.-L.Barabasi, R. Albert, H. Jeong, G. Bianconi, Power law distribution of the World Wide Web, Science 287(2000).2115. 通过和来自各个学科的科学家合作，A.-L.Barabasi 将他的模型迅速推广到各个领域，在过去的十年中掀起了一次复杂网络的研究热潮。各类期刊上出现了满足幂律度分布的大量实证网络数据，在此不详述。

Bernardo A. Huberman, Peter L. T. Pirolli, James E. Pitkow, Rajan M. Lukose. Strong Regularities in World Wide Web Surfing. *Science* 280, 95-97 (1998).

这篇文章介绍了一个用户随机跳转浏览网页，产生类 multiplicative process（一堆代表系统动荡的随机变量相乘的过程）过程的互联网增长模型。作者提出网站拥有的网页数的对数增长率满足对数正态分布，并提出满足这种分布的增长最终会导致整个互联网内网站拥有的网页数呈幂律分布。

延伸：
在Bernardo A. Huberman, Lada A. Adamic Growth dynamics of the World-Wide Web *Natur*e, 401, 131(1999)中，作者举出两大引擎的网页搜索数据，证明其搜索到的

网页确实呈现幂律分布。Bernardo A. Huberman & Lada A. Adamic. Evolutionary Dynamics of the World Wide Web. *Arxiv preprint cond-mat*/9901071(1999)再次解释了这个结果。

multiplicative process 增长过程的另外一个名字叫 Gibrat 律，在生物学、地理学和经济学中都有出现。在地理学中可参考 S. Nordbeck, Geografiska Annaler, Urban allometric growth, *Series B, Human Geography* 53 (1971) 54. 在经济学中可参考 D. Canning, L.A.N. Amaral, Y. Lee, M. Meyer, H.E. Stanley, Scaling the volatility of GDP rates, *Economic Letters*, 60 (1998) 335 以及 Jan Eeckhout. Gibrat's Law for (All) Cities. *the American Economic Review* 94, 1429–1451(2004).

Ramon Ferrer i Cancho & Ricard V. Solé. Least effort and the origins of scaling in human language. *PNAS* 100, 788-791(2003).

这篇论文重新诠释了语言学中的"最省力"（least effort）原则，并在信息论的框架下给出了这个概念的具体数学定义。其实，least effort这个概念，本身就是一个很有潜力的概念，由语言学家G. K.Zipf在*Human behavior and the principle of least effort* (Addison-Wesley, Cambridge, MA, 1949)一书中提出来。虽然Zipf发现了语言学，及其他一些数据中的Zipf律，并指出其可能是least effort的后果，但并没有建立完善的数学模型。本论文中，作者提出了一个Speaker和listeners互相博弈的框架：Speaker倾向于每个字都相同，这样在发音的时候是最省力的（最小熵），listener则希望每个字都不同，这样则最容易识别出每个字的意义（最大熵）。博弈的结果是双方平均分担交流的成本，最后导致存在幂律的语言学结构。

Michael Mitzenmacher. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Math*1, 226-251(2003).

这篇论文谈论了可以产生对数正态分布的模型如 multiplicative process，以及一些可以产生幂律分布的模型如 barabasi 的"优先连接"网络模型，Yule 的"遗传变异"生态学模型和 Simon 的"单词增长演化"语言学模型（这两个模型的思想和优先连接模型的思想基本是一致的，但提出时间要早 50 年），Mandelbrot 的"最小信息成本"语言学模型，Miller 的"随机打字"语言学模型。这篇论文一个独特之处在于将对数正态模型和幂律分布模型联系在一起，指出了前者向后者转化的条件，并提出了一种混合形式（double pareto distribution）；这篇论文的另一个独特之处是引用了大量已有的互联网数据研究的文献，讨论了互联网研究者在网络流量，文件大小，网络连接结构等方面的数据分析中对于对数正态和幂律分布的争论。这篇论文对于从事互联网数据研究的同学尤其有参考价值。

Xavier Gabaix, Parameswaran Gopikrishnan, Vasiliki Plerou & H. Eugene Stanley. A theory of power-law distributions in financial market fluctuations. *Nature* 423, 267-270 (2003).

这篇文章介绍了金融市场中的若干种幂律分布（股票对数收益率、股票交易价格和股票交易量等）及其相互关系，并建立了一个最大化利润的购买模型来推导出幂律分布。在这个模型中，股票交易员试图评估一支股票被低估的程度，并给出一个溢价来向持股者提出一定量的购买要求。该溢价的幅度和购买要求得到满足的时间成反比。同时在市场上，股票被低估的价值随时间减少，因此股票交易员者要抢在股票被低估的价值完全消失前完成买进卖出交易。高溢价可以使交易员尽快买到股票，但利润也随之降低；低溢价虽然保留了大部分利润，但要花掉较长的时间来吸引足够的出售者，因此也冒着股票被低估的价值消失的危险，也是对利润的一种损害。通过最大化整个过程中的总利润，作者推导出了若干个幂律模型。值得一提的是，该模型虽然基于正常的理性交易，但诸如1927-1928经济大萧条前期这样的股市大动荡也在模型的预期范围内，因此该模型有较强的解释力。

M.E.J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46, 323-351 (2005).

这篇论文较详细地介绍了幂律分布的特征、列出了十多种类幂律分布的数据，尤其详细地介绍了幂律分布的数学形式以及它的各种数学处理和变形，如切头、累积分布、rank、最小 x 值，归一化等。在文章的后半部分，介绍了指数、倒数、随机游走、yule 过程、相变与临界态、自组织等多个可以产生幂律的数学模型。

Aaron Clauset, Cosma Rohilla Shalizi & M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review* 51, 661-704 (2009).

这篇论文介绍了如何确认和分析实证数据中的幂律分布。主要谈了使用双对数坐标系下线性回归拟合求幂指数可能产生的问题及其原因；如何使用最大似然估计方法拟合幂指数，以及如何使用 ks 值和 likelihood ratios 来确定幂律分布（作者在 Santa Fe Institute 的网站上公布了可用于拟合和检验的 R 和 Matlab 程序）。
这篇文章对于从事复杂系统实证数据分析的同学尤其有参考价值。

D. Yu. Manin. Mandelbrot's Model for Zipf's Law: Can Mandelbrot's Model Explain Zipf's Law for Language? *Journal of Quantitative Linguistics* 16, 274 -285 (2009).

这篇文章详细介绍了Mandelbrot的“最小信息成本”语言学模型，并做了两处改进。一处是重新解释了信息成本的定义，指出Mandelbrot的为解释公式的数学形式而做出的“单词出现的概率与其长度成负指数关系”的假设不符合语言学的实证数据，提出以“语言使用者”的“心智内存”的假设来代替原有假设，可以更好地理解信息成本的数学形式；另一处是指出Mandelbrot的数学模型中的一个潜在危险：其推导出的Zipf律的指数不是趋近于1，而是趋近于无穷大。作者通过在原有公式的信息成本部分增添一个常数项的办法解决了这个问题。