第 58 卷 第 4 期 2009 年 4 月
1000-3290/2009/58(04)/2127-09

物 理 学 报
ACTA PHYSICA SINICA

Vol.58,No.4,April,2009

基于选择模式的幂律生成机制*

尉伟峰†

(北京交通大学电子信息工程学院,北京 100044)

(2008 年 6 月 27 日收到;2008 年 9 月 8 日收到修改稿)

自然界与社会生活中存在多种性质迥异的幂律分布现象,因而对它们的研究具有广泛而深远的意义.研究了一类由于人类的选择行为而产生的幂律现象,提出了人类选择行为的两个特性:理性与自利.据此,通过小球选择试验构建了模型.研究发现,这类幂律现象中随机变量服从一种近似于 Zipf 分布的分布.此分布为首次提出的一种新分布,将它命名为偏序分布.网络中已有的统计数据表明,偏序分布比 Zipf 分布更真实的反应了网络中统计变量服从的分布.文中对偏序分布进行了理论分析,并通过约束条件的改变,生成了 Yule 分布;表述了 Yule 分布的物理意义,并构造了比 Yule 分布更具现实意义的分布.说明了选择理论在解释幂律现象方面的普遍适用性及模型的可扩展性.文中在幂律现象的成因上支持还原论,即简单的人类选择行为导致了普遍的幂律现象.

关键词:选择行为,幂律分布,偏序分布,Yule 分布

PACC:0250,0500

1. 引 言

幂律分布也称为 Pareto 分布或 Zipf 分布,幂律现象是指系统中某个随机变量统计数据的分布近似服从 Zipf($P_k \sim k^{-r}$)分布的现象.研究发现,幂律现象广泛存在于自然界与社会生活中,尤其是各种结构与功能复杂的网络中[1—9].揭开幂律的成因已成为许多学者关注的一个焦点.

几十年来,为了解释幂律的形成原因,科学家们提出了几种机制,包括增长与优先连接[10,11]、自组织临界[12,13]、最优抗干扰(HOT)理论[14,15]、渗流模型[16—19]以及一些随机过程[16,20,21]等.这些理论分别针对一些特定的幂律现象并合理地解释了其形成原因,而事实上还有相当多的幂律现象没有得到合理地解释.现实世界中存在一类幂律现象,如网页被点击次数的分布、书籍及唱片的销售量的分布等,就本质而言,这些幂律分布是人们的选择行为导致的结果.本文针对这一类幂律现象,研究了人类选择行为的特性,并据此建立数学模型表征了这类幂律现象的生成机制;同时利用选择理论揭示了 Yule 分布在此层面上的物理意义,并分析了选择理论的适用性及其模型的可扩展性.此外,文中还提出了比较函数,对几种不同的分布进行了比较.

2. 人类选择行为的特性及模型的构建

现实生活中,有很多幂律现象的形成是由于人类的选择行为而导致的结果.以网页被点击次数的分布为例,尽管中国向 7900 万网民提供的网站接近 60 万个,但只有为数不多的网站,才拥有网民一次访问难以穷尽的丰富内容,拥有接纳许多人同时访问的足够带宽,进而通过众多网民的选择使其演化成热门网站,拥有极高的点击率,像新浪、搜狐等门户网站.

通过上述案例我们可以看出,网页被点击次数的分布实际上是由于众多网民对网页的选择而产生的结果,而网民对网页的选择也是出于一定的原因的,如网页的内容丰富,带宽的限制等.由此我们发现这类幂律现象的成因同人类的选择行为是密切相关的.

2.1. 人类选择行为的特性分析

人们的任何选择行为,都应该既是合目的性的,也是合规律性的.它是建立在事实和价值统一基础上的满意性.法国启蒙思想家霍尔巴赫说:"利益是

* 国家自然科学基金(批准号:60772043,60672069)和国家重点基础研究发展计划(批准号:2007CB307101)资助的课题.

† E-mail:norax@sina.com

人的行动的唯一动力”.马歇尔和帕累托相互独立地从各自立场出发,都确立了这样的观念:个体以理性的方式行事,以图使自己的个人利益最大化.个体在计算是否投入某项行动时,依据的是该行动能在多大程度上满足自己的需求.

这都非常明确地表述了利益对人们的至关重要性,利益最大化是选择的出发点.因此可以把人们的选择看作是为了实现利益或效用最大化的过程.同时,对于一个理性的个人而言,这里的利益最大化的形式是多种多样的,可以是物质利益最大化,或是精神上的效用最大化,抑或是两者的结合.

由此可以看出人类的选择行为具有以下两个重要的特征:(1)人类的选择行为是理性的,个体独立的、客观的做出选择;(2)人类选择行为的目的是实现自身利益最大化,即人类的选择行为是自利的.

2.2. 模型的构建

以下将通过试验,根据人类选择行为的特征来构建数学模型.

假设存在 N 个大小不同的小球,小球越大则表示其代表的利益越大.将小球由大到小顺序编号,即 1 号小球最大,2 号小球次之,依此类推 N 号小球最小.每个人独立的选择一个小球,记录选择结果并进行统计.

如果每个人的视野不受限制,即其可选择的小球范围是所有的 N 个小球,那么勿容置疑,所有人都会选择 1 号小球.但事实并非如此.在现实生活中,每个人的视野受到各种因素的限制,可供其选择的范围是变化的.因此,人们实现自身利益最大化的选择只是相对的利益最大的选择.也正由于如此,在有些人看来是差的选择,在有些人眼中却是好的选择.由于每个人视野的限制,才呈现出多元化的选择结果.首先考虑最基本的情况,即对每个人而言,所有的视野情况出现的概率相同,每个小球被选到视野中的概率相同.当然这里不存在视野中可选对象为零的情况,即视野中至少存在一个可选对象.

每个人都会选择视野范围内利益最大的小球,所以 k 号小球被取到的前提条件是视野中一定有 k 号小球,其余则为 $k+1$ 号到 N 号这 $N-k$ 个小球中的任意个,因此视野中小球的个数为 i 个,$1 \leqslant i \leqslant N-k+1$;而 1 号到 $k-1$ 号小球则不在视野范围内.如果小球被选进视野中的概率为 p,那么 k 号小球被选到时,所有可能的视野情况可以表示为 $\sum_{i=1}^{N-k+1} C_{N-k}^{i-1} p^i (1-p)^{N-i}$,对 N 个小球选择的统计结果中,k 号小球被选到的概率为:

$$P_k = \frac{\sum_{i=1}^{N-k+1} C_{N-k}^{i-1} p^i (1-p)^{N-i}}{\sum_{k=1}^{N} \sum_{i=1}^{N-k+1} C_{N-k}^{i-1} p^i (1-p)^{N-i}} = \frac{p(1-p)^{k-1}}{1-(1-p)^N}. \tag{1}$$

上述选择过程中,只考虑了小球是否被选进视野中,没有考虑选择小球时的先后次序,即进行视野中小球的选择过程时,只考虑了小球的组合情况,而没有考虑小球的排列情况.那么在选择小球时,将小球的排序考虑进去后,不同大小的小球被选择的概率有何变化呢?

此时,k 号小球被取到的前提条件仍然是视野中一定有 k 号小球,其余则为 $k+1$ 号到 N 号这 $N-k$ 个小球中的任意个,视野中小球的个数依然为 i 个,$1 \leqslant i \leqslant N-k+1$;1 号到 $k-1$ 号小球仍然不在视野范围内.不过由于选择时还需考虑小球的排序,如果小球的排序不影响其被选入视野的概率,那么得到同一个视野的次数为不考虑小球排序时的 $N!$ 倍,即相对于不考虑小球排序的情况,小球被选到时所有可能的视野情况的数量增加了 $N!$ 倍.假定小球被选进视野中的概率为 p,k 号小球被选到时,所有可能的视野情况则为 $\sum_{i=1}^{N-k+1} N!\, C_{N-k}^{i-1} p^i (1-p)^{N-i}$,对 N 个小球选择的统计结果中,k 号小球被选到的概率为:

$$P_k = \frac{\sum_{i=1}^{N-k+1} N!\, C_{N-k}^{i-1} p^i (1-p)^{N-i}}{\sum_{k=1}^{N} \sum_{i=1}^{N-k+1} N!\, C_{N-k}^{i-1} p^i (1-p)^{N-i}} = \frac{\sum_{i=1}^{N-k+1} C_{N-k}^{i-1} p^i (1-p)^{N-i}}{\sum_{k=1}^{N} \sum_{i=1}^{N-k+1} C_{N-k}^{i-1} p^i (1-p)^{N-i}} = \frac{p(1-p)^{k-1}}{1-(1-p)^N}. \tag{2}$$

从(1)式和(2)式可以看出,同选择时只考虑小球的组合相比,如果选择过程中小球的先后排序不影响其被选入视野的概率,即处于选择序列中不同位置的同一个小球是否被选入视野的概率是相同的,此时小球的排序对最终的选择结果并无影响.(1)式和(2)式相同,说明通过这两种选择过程构建

的模型本质上是无差别的.将此模型记为模型 1.

图 1 是模型 1 的仿真与理论结果.从(1)式与图 1 可知,模型 1 的概率分布为几何分布.

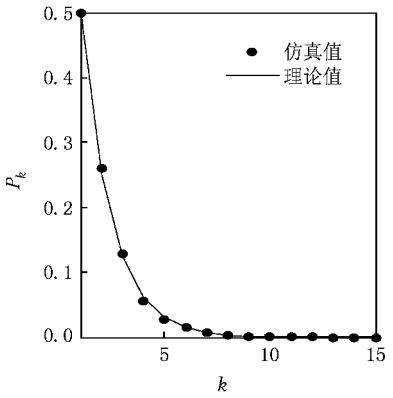

图 1　模型 1 的仿真图与理论图 实验次数为 50 万次,$p=0.5$,$P_k \sim 2^{-k}$

通过上面的分析可知,当小球的排序对其被选入视野不存在影响时,考虑小球的排序与否对模型并无影响.那么当小球的排序对其被选入视野存在影响时,即处于选择序列中不同位置的同一个小球被选入视野的概率不同时,小球的排序会不会对模型产生影响呢?当小球的排序对其被选入视野存在决定性影响时,例如规定只有选择时位于前 m 个的小球能被选入视野,而位于 m 及 m 以后的小球则不能被选入视野,在选择视野过程中,如果 k 号小球的选择顺序为前 m 个,则其必然被选入视野;反之则必然不在视野中.此时 k 号小球所在的位置不同,其被选入视野的概率不同,从而影响了其最终被选中的概率.显而易见,此时小球的排序对最终的选择结果是有影响的,这种情况下构建的模型必然与模型 1 不同.这种情形在现实生活中是很常见的.以网络中的文件下载为例,如果网络中存在 N 个用户,用户 A 可以从这 N 个用户中的任意一个下载所需的文件.由于网速等因素的影响,用户 A 对这 N 个用户的选择存在偏好,假设用户 A 最愿意选择 1 号用户下载文件,2 号次之,依此类推.而这 N 个用户拥有用户 A 所需文件的时间先后次序是等概的.当用户 A 在某时刻 t 需要下载文件 b 时,用户 A 的视野为此时拥有文件 b 的用户.如果 1 号用户此时没有用户 A 所要下载的文件 b,1 号用户必然不在用户 A 的视野中.这里的某时刻 t 相当于一个门限,在此门限范围内,即 t 时刻及 t 时刻前拥有文件 b 的用户被选入视野,否则不能被选入.

可以看出,此时在选择视野的过程中,门限的设置使得处于选择序列中不同位置的同一个小球被选入视野的概率总和不再相同,进而影响了其最终被选到的概率.假设小球被选入视野的概率为 p.当 $p=0.5$ 时,门限的位置是等概产生的,即某时刻在门限允许范围内能被选入视野的小球个数是等概的;当 p 增大时,视野就偏向扩大,反之亦然.此时 k 号小球被选到的条件是 k 号小球的选择顺序必定在门限允许的范围内任意位置,且门限允许范围内的其余小球为 $k+1$ 号到 N 号小球这 $N-k$ 个小球中的任意个小球,因此视野中小球的个数为 i 个,$1 \leqslant i \leqslant N-k+1$;而 1 号小球到 $k-1$ 号小球不管其如何排列,其必定位于门限允许的范围之外.k 号小球被选到时,所有可能的视野情况可以用数学表达式表示为 $\sum_{i=1}^{N-k+1} C_{N-k}^{i-1} i\,!(N-i)\,!p^i(1-p)^{N-i}$,对 N 个小球选择的统计结果中,k 号小球被选到的概率为:

$$
\begin{aligned}
P_k &= \frac{\sum_{i=1}^{N-k+1} C_{N-k}^{i-1} i\,!(N-i)\,!p^i(1-p)^{N-i}}{\sum_{k=1}^{N}\sum_{i=1}^{N-k+1} C_{N-k}^{i-1} i\,!(N-i)\,!p^i(1-p)^{N-i}} \\
&= \frac{N!\sum_{i=1}^{N-k+1} C_{N-k}^{i-1}/C_N^i p^i(1-p)^{N-i}}{N!\sum_{k=1}^{N}\sum_{i=1}^{N-k+1} C_{N-k}^{i-1}/C_N^i p^i(1-p)^{N-i}} \\
&= \frac{\sum_{i=1}^{N-k+1} C_{N-k}^{i-1}/C_N^i p^i(1-p)^{N-i}}{\sum_{k=1}^{N}\sum_{i=1}^{N-k+1} C_{N-k}^{i-1}/C_N^i p^i(1-p)^{N-i}}.
\end{aligned}
\tag{3}
$$

(3)式与(1)式是截然不同的,将(3)式代表的模型记为模型 2.当 $p=0.5$,图 2 是模型 2 的仿真与理论结果,在对数图中,它的曲线近似为直线.这说明人类的选择行为导致了幂律现象的产生.当 p 取不同的值时,各分布见图 3.改变 p 的取值,即可以产生不同幂指数的幂律分布.

当 $p=0.5$,根据组合公式(推导过程见附录 A),(3)式可表示如下.

$$
P_k = \frac{N+1}{N}\frac{1}{k(k+1)}, \tag{4}
$$

当 N 和 k 远远大于 1 时,有 $P_k \sim k^{-2}$.

公式(3)阐述了人类的选择行为特征.当 k 较大时其分布近似于 Zipf 分布.我们将其称为偏序分布.

对数图上偏序分布曲线的头部和尾部是不一样的,整体呈现出两段特性,这符合大量的网络测量数

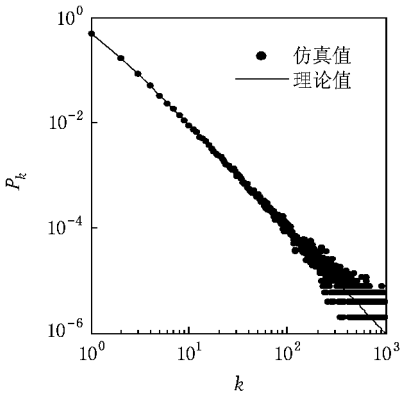

图 2 模型 2 的仿真图与理论图实验次数为 50 万次，$p=0.5$，$P_k \sim k^{-2}$

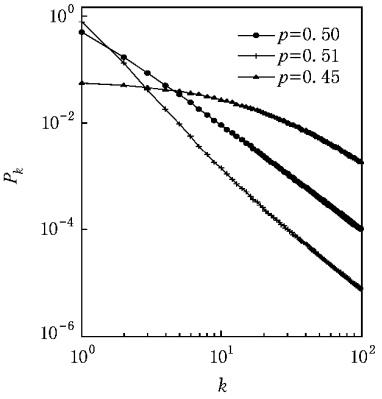

图 3 当 $N=100$，p 分别为 0.45，0.50，0.51 时，模型 2 的分布曲线图

据.当 N 远大于 1 时，观察 P_{N-1} 和 P_N，推导可得偏序分布尾部的幂指数 r 和概率 p 的关系如下，

$$r = \frac{2p}{1-p}.$$

可见，当 p 值介于 0.5 与 0.6 之间时，即得到指数 r 介于 2 与 3 之间的幂律分布，而现实生活中存在的幂律分布其幂指数大多在 2 到 3 的范围内.当 N 趋向无穷时，观察 P_1 和 P_2 可知，$p>1/2$ 时，头部会越来越陡，直到斜率趋向无穷；而 $p<1/2$ 时，头部会越来越平，直到斜率趋向零；而当 $p=1/2$ 时，在对数图上，头部通过幂指数为 $r=\log_2 3$ 的直线.

当 N 较大时，与超几何函数展开式对比，公式(3)可以近似表示为如下形式：

$$P_k \doteq \frac{{}_2F_1(2,k-N;1-N;p/(1-p))}{\sum_{k=1}^{N} {}_2F_1(2,k-N;1-N;p/(1-p))},$$

其中函数 ${}_2F_1(.)$ 为第一类超几何函数，可知偏序分布具有超几何特性.

从上述两个模型的构建过程中可以看出，模型 2 是在模型 1 的基础上，加入了限制条件而构建的，因此模型 2 与模型 1 既有联系，同时又存在一定的区别.

模型 1 和模型 2 最大的不同之处在于，它们的视野范围形成的方式不同.在模型 1 中，同一个小球处于选择序列中的不同位置时，它被选入视野的机会是相同的；但在模型 2 中由于门限的存在，同一个小球处于选择序列中的不同位置时，它能被选入视野的机会是不同的.可以认为模型 1 和模型 2 是反映人类选择行为特征的两个最基本的模型.

以 $p=0.5$ 为例，模型 1 的均值和方差近似如下(推导过程见附录 A.2)，

$$E(L) \approx \frac{N}{2}, D(L) \approx \frac{N}{4} N \gg 1.$$

模型 2 的均值和方差如下，

$$E(L) = \frac{N+1}{2}, D(L) = \frac{N^2-1}{12}.$$

其中，N 是被选对象的总数，通常数值较大，L 是视野中可选对象的个数，为随机变量.模型 2 的方差大于模型 1，因此模型 2 的概率分布更为分散，尾巴更长[22].

当 N 足够大时，两个模型的均值基本相同，但是它们的可选对象集合长度的方差不同.当 $p=0.5$ 时，模型 1 中可选对象集合的长度服从二项分布(视野个数为零的除外)，模型 2 中可选对象集合的长度服从等概分布.

3. 理论分析

上节从数学理论的角度建立了相应的模型，本节将重点讨论模型的现实意义及模型的扩展性.此外，本节提出了比较函数，对几种不同的分布进行了比较.

3.1. 模型的现实意义

现实生活中，由于各种因素的限制，可供人们选择的范围往往不是所有的可选对象.因此，模型 2 比模型 1 更贴近现实.以下将利用模型 2 来解释网络中文件下载次数的分布.网络中，用户对从其他用户下载文件存在一个偏好顺序.用户产生这种偏好的原因很多，如网速快、质量好等.假设网络中存在用户 1，用户 2，一直到用户 10 共 10 个用户.以用户 3

作为观察对象，用户 3 最希望从用户 4 下载文件，依此类推，形成了用户 3 对从其他 9 个用户下载文件的偏好顺序．一般来说，网络中的用户下载文件 a 的时刻是等概随机的．不考虑用户下载文件耗费的时间，用户 3 在时刻 t 下载文件 a 时，首先去看此时刻前，网络中哪些用户已经拥有了文件 a，然后再从这些用户当中挑一个相对最喜欢的用户下载文件．在这里，用户 3 下载文件 a 的时刻 t 相当于模型 2 中的门限．用户下载不同文件的等概随机顺序构成了模型 2 中选择小球的序列；观察对象用户 3 下载文件时刻的等概性对应了模型 2 中门限位置的等概性．图 4 描述了用户 3 下载文件的过程．用户 3 从其余用户下载文件的过程是同模型 2 相对应的．因此，网络中某一用户从其他用户下载文件次数的分布服从偏序分布，我们把这种带有偏好控制的文件生成树称为偏序树，以区别随机生成树．

网络中其余由于人类的选择行为形成的幂律现象也都可以用模型 2 来解释．图 5 是网络中网站数与用户接入数排序关系统计数据图[23]．比较图 5 与图 2 可以发现，模型 2 的仿真图与现实的统计数据图基本一致．这说明用模型 2 解释网络中由于人类选择行为形成的幂律现象的合理性．

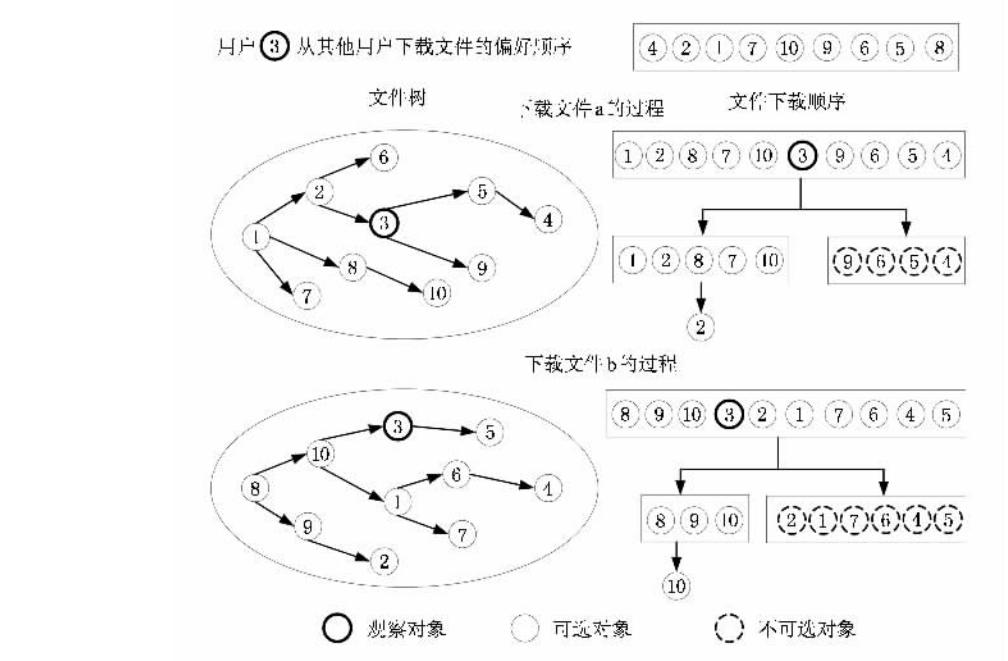

图 4　网络中用户下载文件过程示意图

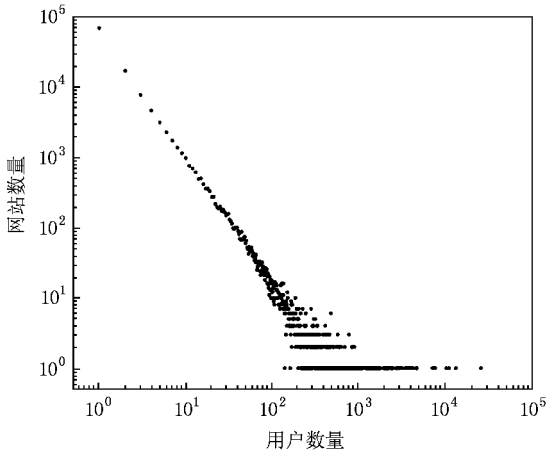

图 5　网络中网站数与用户数关系统计数据图

3.2. 模型的扩展

Yule 分布的形成可以利用人类选择行为特性理论来解释．模型 2 中，每个人最终只能选择一个小球，如果每个人最终选择的小球为序号连续的 r 个小球时，$1 \leqslant r \leqslant N$，统计结果又如何变化呢？此时将最终选择结果中选到序号相邻的 k 号小球，$k+1$ 号小球，一直到 $k+r-1$ 号小球共 r 个小球的情况作为 k 号小球被选中的情况．k 号小球被选到时，视野中除了必须有 k 号小球到 $k+r-1$ 号小球外，还可以有序号为 $k+r$ 号以后的 $N-k-r+1$ 个小球中的任意 j 个，$0 \leqslant j \leqslant N-k-r+1$，1 号小球到 $k-1$ 号小球则不能在视野中出现，因此所有可能的视野情况可以用数学表达式表示为

$$\sum_{i=1}^{N-k-r+2} C_{N-k-r+1}^{i-1}(i+r-1)!\,(N-i-r+1)!\,p^{i+r-1}(1-p)^{N-i-r+1},$$

对 N 个小球选择的统计结果中，k 号小球被选到的概率为：

$$P_k = \frac{\sum_{i=1}^{N-k-r+2} C_{N-k-r+1}^{i-1}/C_N^{i+r-1}p^{i+r-1}(1-p)^{N-i-r+1}}{\sum_{k=1}^{N-r+1}\sum_{i=1}^{N-k-r+2} C_{N-k-r+1}^{i-1}/C_N^{i+r-1}p^{i+r-1}(1-p)^{N-i-r+1}} \quad (k=1,2,\cdots,N-r+1). \tag{5}$$

当 $p=0.5$ 时，(5)式可简化为(推导过程见附录 C)

$$P_k \sim \frac{1}{(k+r)(k+r-1)\cdots k} = \frac{(k-1)!}{(k+r)!}, \tag{6}$$

当 N 趋于无穷大，$p=0.5$，推广 r 为实数时，$P_k = rB(r+1, k)$ $(k=1,2,\cdots,\infty)$，这就是 Yule 分布，其中 $B(r+1, k)$ 是贝塔函数.

由此我们利用人类选择行为特性理论建立模型生成了 Yule 分布. 从这个意义上而言，Yule 分布反映的是选择序号相邻的 r 个元素的情况. 如当 $r=2$，仍然以小球试验为例，此时每个人最终可以选择序号相邻的 2 个小球. 将序号相邻的两个小球看作 1 个最终选择元素，如将 1 号小球和 2 号小球的组合(1, 2)或(2, 1)看作新的 1 号元素，(2, 3)或(3, 2)为 2 号元素，依此类推，$(N-1, N)$ 或 $(N, N-1)$ 为 $N-1$ 号元素，而譬如(1, 3)，(3, 5)等不相邻的 2 个小球组合被选到的概率则不予考虑. 模拟仿真结果显示，1 号元素被选到的次数最多，2 号元素被选中的次数次之，依此类推，$N-1$ 号元素被选到的次数最少；同时 1 号元素到 $N-1$ 号元素被选到的概率分布服从 Yule 分布. 将 $r=2$ 的情况进行推广，可以得出在序号相邻的 r 个小球被选到的情况下，新产生的 1 号到 $N-r+1$ 号元素被选中的概率分布服从 Yule 分布.

不难看出，除了 $r=1$ 的情况外，Yule 分布所反应的物理意义在现实生活中发生的可能性不大. 从此意义上而言，Yule 分布并不具有很强的现实意义.

如果最终允许选择的小球个数为 2，那么更符合实际的选择情况的是人们从视野中选出 2 个最好的. 据此对模型 2 进行了扩展. 此时 k 号小球被选到时，视野中除了必须有 k 号小球外，还可以存在的其余小球可分为两部分，一部分为 $k+1$ 号小球到 N 号小球这 $N-k$ 个小球，另一部分为 1 号小球到 $k-1$ 号小球中的任意一个. 即此时的视野情况可分为两种类型，一类为视野中除了 k 号小球外，其余小球为 $k+1$ 号小球到 N 号小球这 $N-k$ 个小球中的任意 j 个，$0 \leqslant j \leqslant N-k$，其数学表达式为 $\sum_{i=2}^{N-k+1} C_{N-k}^{i-1} i\,!(N-i)!\,p^i(1-p)^{N-i}$；另外一类为视野中除了 k 号小球外，其余小球为 1 号小球到 $k-1$ 号小球中的任意一个和 $k+1$ 号小球到 N 号小球这 $N-k$ 个小球中的任意 j 个，$0 \leqslant j \leqslant N-k$，其数学表达式为 $\sum_{i=2}^{N-k+2} C_{N-k}^{i-2}C_{k-1}^{1}\, i\,!(N-i)!\,p^i(1-p)^{N-i}$. 因此 k 号小球被选到时，所有可能的视野情况为 $\sum_{i=2}^{N-k+1} C_{N-k}^{i-1} i\,!(N-i)!\,p^i(1-p)^{N-i} + \sum_{i=2}^{N-k+2} C_{N-k}^{i-2}C_{k-1}^{1}\, i\,!(N-i)!\,p^i(1-p)^{N-i}$，对 N 个小球选择的统计结果中，k 号小球被选到的概率为：

$$P_k = \frac{\sum_{i=2}^{N-k+1} C_{N-k}^{i-1}/C_N^i p^i(1-p)^{N-i} + \sum_{i=2}^{N-k+2} C_{N-k}^{i-2}C_{k-1}^{1}/C_N^i p^i(1-p)^{N-i}}{\sum_{k=1}^{N}\left(\sum_{i=2}^{N-k+1} C_{N-k}^{i-1}/C_N^i p^i(1-p)^{N-i} + \sum_{i=2}^{N-k+2} C_{N-k}^{i-2}C_{k-1}^{1}/C_N^i p^i(1-p)^{N-i}\right)} \quad (k=2,3,\cdots,N),$$

$$P_1 = P_2. \tag{7}$$

当 $p=0.5$ 时，式(7)可以化简为：

$$P_k \sim \sum_{i=2}^{N-k+1} C_{N-k}^{i-1}/C_N^i + \sum_{i=2}^{N-k+2} C_{N-k}^{i-2}C_{k-1}^{1}/C_N^i = \frac{3(N+1)}{k(k+1)} - \frac{1}{N} \quad (k=2,3,\cdots,N),$$

$$P_1 = P_2. \tag{8}$$

此模型的仿真结果见图 6. 图中，随机变量概率分布的幂指数近似为 2.

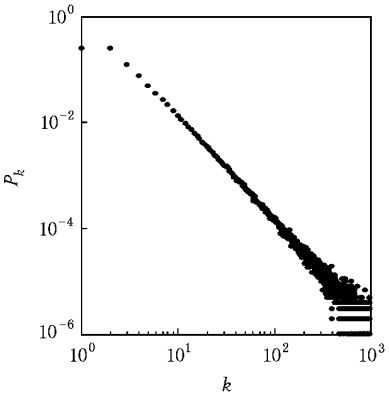

图 6　模型的仿真图

此外，还可以从其他方面对模型 2 进行扩展. 如模型中每个小球被不同的人选择的概率可以不相同；小球的排序对不同的人而言可以不同等等. 不难看出，利用人类选择行为特性理论建立的模型具有很强的扩展性，通过改变各种约束条件，可以生成物理意义不同的各种分布. 因此，通过此理论建立的模型更具普遍的现实意义.

3.3. 分布的比较

为了比较不同分布之间的区别，提出了比较函数 $\zeta(l)$. $\zeta(l)$ 为不同分布的相邻两个等级的元素被选到的概率之比，这里的等级即为上述模型中小球的大小，小球越大，等级越高. 表 1 是各分布的 $\zeta(l)$.

$$\zeta(l) \triangleq \frac{P_l}{P_{l+1}}.$$

表 1　各分布的 $\zeta(l)$

分布类型	等概分布	几何分布	Yule 分布	幂律分布
	$P_k \sim \text{const.}$	$P_k \sim r^{-k}(r \neq 1)$	$P_k \sim (k-1)!/(k+r)!$	$P_k \sim k^{-r}$
函数 $\zeta(l)$	1	r	$1+(r+1)/l$	$(1+1/l)^r$

以选择小球试验为例，函数 $\zeta(l)$ 反映了 l 号小球和 $l+1$ 号小球被选到的概率相差多少倍. 从表 1 中可以看出，等概分布中，不同等级的元素被选到的概率相同；而几何分布、Yule 分布和 Zipf 分布中，不同等级的元素被选到的概率是不同的，且等级越高的元素被选中的概率越高；且 Yule 分布中，选择行为呈现双曲线特征. 可以看出，不同的分布中，其不同等级元素被选中的概率的差异性是不同的，$\zeta(l)$ 函数为比较不同的分布提供了一个平台. 特别在分析数据时，也可以先做出 $\zeta(l)$ 的拟合曲线，再用递推法算出其解析分布.

4. 结　论

幂律分布已有超过 100 年的研究历史了，即使在现在，仍然是众多学科研究的热点. 统计物理学家习惯于把服从幂律分布的现象称为无标度现象. 在自然界和社会生活中，凡有生命的地方，有进化、有竞争的地方都会出现不同程度的无标度现象. 通常而言，有进化、有竞争的地方意味着选择行为的存在. 从这个意义上而言，利用选择行为理论构建模型解释幂律现象具有一定的普遍的现实意义. 此外文中所提出的文件下载模型也可以用来构建幂律传播模型，适合于研究病毒、谣言等传播机制.

在研究幂律现象生成机制的过程中，本文中首次提出了一种新的分布——偏序分布. 偏序分布的生成机制是很简单的，它根据选择行为理论，利用概率论构建模型得到，且易于扩展. 而简单性一向是现代自然科学、特别是物理学的一条重要的指导原则[24]. 偏序分布是一个很普遍的分布，网络中及人类社会系统中的很多幂律现象都可以用偏序试验来解释. 虽然网络与人类的行为是一个复杂的过程，但是从心理和物理的角度来看，他们都是在随机物理条件下功利主义的结果. 因此我们可以直接深入研究这个规律，而不关注其复杂的内部机制. 本文在幂律现象的成因上支持还原论，即简单的人类选择行为导致了普遍的幂律现象.

衷心感谢陈常嘉教授、郭宇春博士、张敏博士、郭海燕硕士对本文工作提出的建议.

附录 A　公式(4)的推导

由组合公式可以得到：

$$\frac{C_{N-k}^{i-1}}{C_N^i} = \frac{i!(N-k)!(N-i)!}{(i-1)!(N-k-i+1)!N!}$$
$$= \frac{i(N-i)!}{k(k-1)!(N-i-k+1)!}\frac{k!(N-k)!}{N!}$$
$$= \frac{iC_{N-i}^{i-1}}{kC_N^k}.$$

因此有：

$$\sum_{i=1}^{N-k+1}\frac{C_{N-k}^{i-1}}{C_N^i} = \frac{1}{kC_N^k}\sum_{i=1}^{N-k+1} iC_{N-i}^{k-1},$$

推导有恒等式为：

$$\sum_{i=1}^{N-k+1} i(x+1)^{N-i} \equiv -\frac{[(N-k+2)x+1](x+1)^{k-1}}{x^2} + \frac{(x+1)^{N+1}}{x^2},$$

比较等式两边 x^{k-1} 的系数,即得到

$$\sum_{i=1}^{N-k+1} iC_{N-i}^{k-1} = C_{N+1}^{k+1},$$

于是可得

$$\sum_{i=1}^{N-k+1} \frac{C_{N-k}^{i-1}}{C_N^i} = \frac{C_{N+1}^{k+1}}{kC_N^k} = \frac{N+1}{k(k+1)},$$

$$\sum_{k=1}^{N}\sum_{i=1}^{N-k+1} \frac{C_{N-k}^{i-1}}{C_N^i} = (N+1)\sum_{k=1}^{N}\left(\frac{1}{k}-\frac{1}{k+1}\right) = N.$$

因此

$$P_k = \frac{N+1}{N}\frac{1}{k(k+1)} \qquad k = 1,2,\cdots,N,$$

附录 B 模型 1 均值和方差的推导结

$$E(L) = \sum_{k=1}^{N} k\frac{C_N^k}{2^N-1} = \frac{N}{2}\frac{2^N}{2^N-1},$$

$$D(L) = \sum_{k=1}^{N}(E(L)-k)^2\frac{C_N^k}{2^N-1} = \frac{N(-2\cdot 4^N-4^N N+8^N+2^N+2^N N)}{4\cdot(2^N-1)^3}.$$

附录 C 公式(6)的推导,其过程与附录 A 类似

$$\sum_{i=1}^{N-k-r+2} \frac{C_{N-k-r+1}^{i-1}}{C_N^{i+r-1}} = \frac{N+1}{(k+r)C_{k+r-1}^r},$$

$$\sum_{k=1}^{N-r+1}\sum_{i=1}^{N-k-r+2} \frac{C_{N-k-r+1}^{i-1}}{C_N^{i+r-1}} = (N+1)\cdot\frac{C_{N+1}^r-1}{rC_{N+1}^r},$$

$$P_k = \frac{rC_{N+1}^r}{C_{N+1}^r-1}\cdot\frac{1}{(k+r)C_{k+r-1}^r} \sim \frac{(k-1)!}{(k+r)!} \qquad k = 1,2,\cdots,N-r+1.$$

[1] Zipf G K 1949 *Human Behavior and the Principle of Least Effort* (Cambridge: Addison-Wesley) p1

[2] Krapivsky P L, Redner S 2001 *Phys. Rev.* E **63** 066123

[3] Barabási A L, 2002 *Linked: The New Science of Networks* (Cambridge: Perseus) p1

[4] Albert R, Barabási A L 2002 *Rev. Mod. Phys.* **74** 47

[5] Dorogovtsev S N, Mendes J F F 2003 *Evolution of Networks——From Biological Nets to the Internet and WWW* (Oxford: Oxford University Press) p1

[6] Satorras R P, Vázquez A, Vespignani A 2001 *Phys. Rev. Lett.* **87** 258701

[7] Vázquez A, Satorras R P, Vespignani A 2002 *Phys. Rev.* E **65** 66130

[8] Satorras R P, Vespignani A 2004 *Evolution and Structure of the Internet-A Statistical Physics Approach* (Cambridge: Cambridge University Press) p1

[9] Hu H B, Wang L 2005 *Physics* **34** 889 (in Chinese) [胡海波、王林 2005 物理 **34** 889]

[10] Albert R, Barabási A L 1999 *Science* **286** 509

[11] Barabási A L, Albert A 1999 *Physica* A **272** 173

[12] Bak P, Tang C, Wiesenfeld K 1987 *Phys. Rev. Lett.* **59** 381

[13] Bak P 2001 *How Nature Works* (Wuhan: Central China Normal University Press) p10 (in Chinese) [帕巴克 2001 大自然如何工作(武汉:华中师范大学出版社)第 10 页]

[14] Carlson J M, Doyle J 1999 *Phys. Rev.* E **60** 1412

[15] Carlson J M, Doyle J 2000 *Phys. Rev. Lett.* **84** 2529

[16] Newman M E J arXiv: cond-mat/0412004 v3 [cond-mat.stat-mech]

[17] Broadbent S R, Hammersley J M 1957 *Proc. Cambridge Philos. Soc.* **53** 629

[18] Hammersley J M 1957 *Proc. Cambridge Philos. Soc.* **53** 642

[19] Grimmett G 1999 *Percolation* (2nd ed) (Berlin: Springer-Verlag) p1

[20] Reed W J, Hughes B D 2002 *Phys. Rev.* E **66** 067103

[21] Mitzenmacher M 2004 *Internet Math* **1** 226

[22] Willinger W, Anderson D, John C D, Li L 2004 *Proc.* 2004 *Winter Simulation Conf.* Washington DC, December, 2004 (New York: IEEE Press) p130

[23] Adamic L A, Huberman B A 1999 *The Nature of Markets in the World Wide Web* (Palo Alto: Xerox Palo Alto Research Center) p1

[24] Hao B L 2001 *Physics* **30** 466 (in Chinese) [郝伯林 2001 物理 **30** 466]

Power law phenomena based on human behaviors*

Wei Wei-Feng†

(*School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing* 100044, *China*)

(Received 27 June 2008; revised manuscript received 8 September 2008)

Abstract

Various phenomena governed by power law distributions are common in nature and in society, thus their study has broad and far-reaching significance. This paper focuses on a kind of power law phenomenon originating from human behaviors, for instance, popular web page or blog distribution and so on. It is pointed out that human behaviors are rational and self-interested. On the basis of this, we established a model and found that variables of this kind obeyed a new distribution which was named as the Wei distribution rather than Zipf distribution. Based on Wei distribution, Yule distribution was generated through changing the restrictions. Moreover, in the paper, we describe the meaning of Yule distribution and set up a new distribution which would characterize the real-life conditions more faithfully. It is illustrated further that the human behavior theory could be applied generally to explaining this kind of power law phenomenon and the model could be expanded easily. Besides, we put forward a function to compare different kinds of distributions. This paper supports the reductionism in power law phenomena, which means that simple human behaviors lead power law to be universal.

Keywords: human's behaviors, power law distribution, Wei distribution, Yule distribution

PACC: 0250, 0500

* Project supported by the National Natural Science Foundation of China (Grant Nos. 60772043, 60672069) and the State Key Development Program for Basic Research of China (Grant No. 2007CB307101).

† E-mail: norax@sina.com