УДК 519.234.7

**В.Л. Шергин, Э.Э.Дереза, В.С.Передерий, М.Р.Полиит**
ХНУРЭ, г.Харьков, Украина, vadim.shergin@nure.ua

# ОЦЕНИВАНИЕ ПАРАМЕТРА РАСПРЕДЕЛЕНИЯ ЮЛА

Проведён анализ области применения модели Юла и методов оценивания параметра распределения Юла-Саймона. Предложены методы оценивания искомого параметра. Проведён численный эксперимент, в котором моделировалась сеть Барабаши-Альберт и проводился сравнительный анализ точности оценок параметра распределения Юла. Показано, что оценки, полученные на основе ряда отношений обычной функции вероятности к кумулятивной являются несмещёнными.

РАСПРЕДЕЛЕНИЕ ЮЛА, СТЕПЕННОЕ РАСПРЕДЕЛЕНИЕ, СЕТЬ БАРАБАШИ-АЛЬБЕРТ, НЕСМЕЩЁННЫЕ ОЦЕНКИ

## Введение

Одним из самых убедительных и широко применимых механизмов, порождающих степенные законы, является процесс Юла [1]. Он описывает динамику растущей системы, в которой действует принцип *преференциального присоединения* (он же принцип Гибрата, эффект Матфея, «богатый становится богаче»): вероятность присоединения нового микрообъекта к макрообъекту прямо пропорциональна количеству микрообъектов, уже имеющихся у этого макрообъекта.

Изначально сам Юл построил эту модель для описания динамики возникновения новых родов растений (макрообъектов) и объяснения степенного закона их распределения по числу видов (микрообъектов). Закон распределения Юла (называемый также законом Юла-Саймона) имеет вид [2]:

$$p_k = (\theta-1)\cdot \mathrm{B}(k,\theta) = (\theta-1)\frac{\Gamma(k)\Gamma(\theta)}{\Gamma(k+\theta)}, \qquad (1)$$

где $\Gamma(x)$ – гамма-функция, $\mathrm{B}(x,y)$ – бета-функция (эйлеров интеграл первого рода).

Поскольку бета-функция имеет степенной хвост ($\mathrm{B}(k,\theta)\sim k^{-\theta}$ при $k\gg 1$ и фиксированном $\theta$), то распределение Юла (1) асимптотически стремится к (дискретному) степенному закону:

$$p_k \approx C\cdot k^{-\theta}. \qquad (2)$$

Одной из наиболее изученных вычислительных моделей, динамика которой описывается процессом Юла, является модель сети Барабаши-Альберт [3]. Эта модель описывает рост сети (неориентированного графа), при условии, что вероятность присоединения новой вершины к вершине, имеющей степень $k$, прямо пропорциональна значению $k$. В частности, если связи образуются только с новыми вершинами (и тогда граф становится деревом), то распределение вершин по степеням соответствует закону Юла с параметром $\theta = 3$.

## 1. Актуальность проблемы и постановка задачи исследования

Процесс Юла - правдоподобный и достаточно общий механизм, который может объяснить степенные распределения, наблюдающиеся в природе, и при подходящем наборе параметров может генерировать распределения с показателями в большом диапазоне.

Модель Юла была успешно адаптирована и обобщена исследователями для объяснения степенных законов во многих системах [4]. Наиболее известные примеры – распределения городов по числу жителей, домохозяйств по доходам, число цитирований научных статей, число ссылок на веб-страницы всемирной сети [1]. Для перечисленных примеров процесс Юла стал наиболее популярной теорией.

При этом важно отметить, что во всех перечисленных исследованиях модель Юла использовалась именно для объяснения степенных законов распределений, но не для анализа данных. Это было продиктовано простотой анализа степенных распределений и может быть вполне обоснованно при больших количествах макрообъектов (домохозяйств, статей) и неопределённости их нижних границ (величины доходов домохозяйств, населения городов) в исследуемых выборках.

В то же время, при малых значениях $k$ (т.е. численности микрообъектов в макрообъекте), разница между распределениями Юла и степенным является весьма заметной. Так, модель древовидной сети Барабаши-Альберт [3] предполагает $\theta = 3$. В то же время, иногда постулируется [2, 5], что параметр этой модели ($\theta$) может варьироваться от 2 до 3. При этом зачастую происходит попытка компенсировать систематическую ошибку оценивания параметра, вызванную подменой закона распределения, за счёт введения дополнительных параметров $k_0, c$ [4].

Единственным известным методом оценивания параметра распределения Юла ($\theta$), не основанным на сведении закона Юла (1) к степенному, является метод, предложенный Гарсиа [6]. Он представляет собой итеративную реализацию метода наибольшего правдоподобия для закона (1). Несмотря на высокую точность оценивания, этот метод обладает существенным недостатком: сложностью реализации, обусловленной как использованием неэлементарных функций, так и итеративностью метода.

Таким образом, проблема оценивания параметра распределения Юла ($\theta$) является актуальной научной и практической задачей. Для её решения

в настоящей работе рассматриваются и решаются следующие задачи:

– анализ традиционных методов оценивания параметра распределения Юла (на основе log-log регрессии), выявление причин возникновения системной погрешности;

– модификация традиционного метода оценивания $\theta$ путём сведения закона Юла к сдвинутому степенному распределению;

– разработка метода оценивания $\theta$ без использования итераций и степенных распределений;

– сравнение точности полученных оценок с помощью численного моделирования.

## 2. Оценивание параметра распределения Юла с помощью степенного распределения

Как было отмечено во введении, закон распределения Юла (1) при больших значениях $k$ асимптотически стремится к дискретному степенному закону. Поэтому если данные действительно соответствуют закону (1), то зависимость $p_k(k)$ в двойных логарифмических координатах будет очень близка к линейной.

Согласно [7], при практическом анализе безмасштабных данных предпочтительнее пользоваться эмпирическими значениями не самой функции вероятности $p_k$, а (также эмпирическими) значениями дополненной функции распределения (кумулятивной функции вероятности) $G_k = P(x \ge k)$. Для распределения Юла (1) эта функция имеет вид

$$G_k = (\theta - 1) \cdot \mathrm{B}(k, \theta - 1) = \frac{\Gamma(k)\Gamma(\theta)}{\Gamma(k + \theta - 1)}. \qquad (3)$$

При $k \gg 1$ справедливо приближение

$$G_k \approx C \cdot k^{-(\theta - 1)}. \qquad (4)$$

Традиционный метод оценки параметра степенного распределения состоит в применении метода наименьших квадратов (МНК) к прологарифмированной модели (4):

$$\log(G_{k_i}) = \log(C) - (\hat{\theta}^{(1)} - 1) \cdot \log(k_i). \qquad (5)$$

К сожалению, если истинным распределением случайной величины является закон Юла (3), а не степенной (4), то получаемые оценки являются смещёнными. В самом деле, при $k \to \infty$ наклон кривой (5) действительно равен

$$\frac{\log(\mathrm{G}_{k+1}) - \log(\mathrm{G}_k)}{\log(k+1) - \log(k)} = \frac{\log(\frac{k}{k+\theta-1})}{\log(\frac{k+1}{k})} \approx -(\theta - 1).$$

В то же время, при $k = 1$ этот наклон составит $\frac{-\log(\theta)}{\log(2)} = -\log_2 \theta$. На графике (рис. 1) приведены результаты оценивания параметра распределения степеней вершин сети Барабаши-Альберт (построенной согласно процессу Юла с истинным значением параметра $\theta = 3$).

Как и было показано выше, при больших значениях $k$ график дискретного степенного распределения (построенный при $\theta = 3$) практически сливается с графиком распределения Юла, а при малых $k$ заметно отклоняется от него. В результате угол наклона регрессионной прямой (показанной пунктиром) смещается в сторону уменьшения оценки $\theta$.

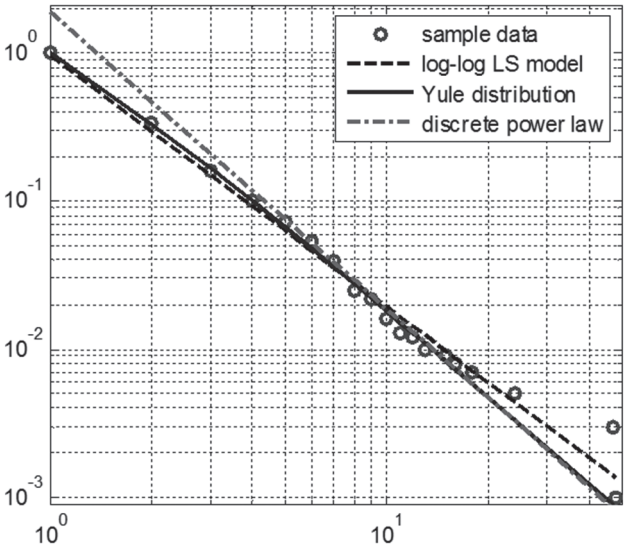


**Рис. 1. Кумулятивная функция вероятности распределения степеней вершин сети Барабаши-Альберт**

Данные численного эксперимента наглядно иллюстрируют, что оценка параметра распределения Юла, полученная традиционным методом сведения его к степенному, является смещённой. Рассчитать математическое ожидание смещения теоретическими методами не представляется возможным. В то же время, получить численную оценку смещений не составляет труда (рис.2).

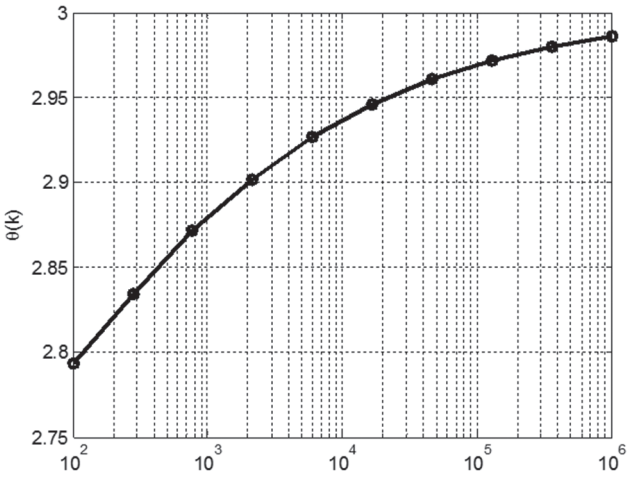


**Рис. 2. Зависимость МНК-оценок параметра распределения Юла от объёма выборки**

Согласно представленной зависимости МНК-оценок параметра $\theta$ от объёма выборки, соответствующей закону Юла при $\theta = 3$, для выборки из 1000 элементов смещение составляет $\Delta\theta(1000) = 0.1208$. Прибавив это значение к $\hat{\theta}^{(1)}$ (рис. 1), улучшаем оценку: $\hat{\theta}^{(2)} = \hat{\theta}^{(1)} + \Delta\theta = 1.6958 + 0.1208 = 1.8166$.

### 3. Оценивание параметра распределения Юла на основе сдвинутого степенного распределения

Без нарушения общности, сравним вид закона Юла (1) и приближающего его степенного закона (2) при $\theta = 3$. В первом случае функция вероятности имеет вид $p_k = \dfrac{4}{k(k+1)(k+2)}$, а во втором – $p_k = C \cdot k^{-3}$. Неудивительно, что наибольшие различия между этими законами наблюдаются при малых значениях $k$.

Представляется логичным искать приближение закона Юла не в форме (2), а в форме

$$p_k = C \cdot (k+m)^{-\theta}. \qquad (6)$$

Константу $C$ находим из условий нормировки закона (6): $C = 1/\zeta(\theta, m+1)$, (где $\zeta(z,v)$ – дзета-функция Гурвица), а константу $m$ – из условия совпадений значений функций вероятности (1) и (6) при $k = 1$:

$$p_1 = \frac{\theta - 1}{\theta} = \frac{(m+1)^{-\theta}}{\zeta(\theta, m+1)}. \qquad (7)$$

Численное решение уравнения (7) показывает, что при $\theta = 3$ $m \approx 0.7369$, а $m(\theta = 2) \approx 0.3896$; зависимость $m(\theta)$ очень близка к линейной в широком диапазоне значений $\theta > 1$. Так, для диапазона $\theta \in [2; 3]$ $m \approx 0.34713 \cdot \theta - 0.30212$. Найденное соотношение позволяет построить итеративную процедуру оценивания параметра распределения Юла: для некоторого начального приближения $m = m_0$ строим модель

$$\log(p_{k_i}) = \log(C) - \hat{\theta}^{(3)} \cdot \log(k_i + m), \qquad (8)$$

по которой находим оценку $\theta$ с помощью МНК. После чего можно уточнить значение параметра $m$ и повторить итерацию.

Вместе с тем, для большинства процессов и явлений, динамика которых объясняется процессом Юла, диапазон значений $\theta$ составляет $\theta \in [2; 3]$, поэтому можно утверждать, что любое значение $m$, взятое из соответствующего диапазона $m \in [0.3896; 0.7369]$, будет являться заведомо лучшим приближением, чем нулевое (соответствующее обычному степенному распределению и модели (2)).

В настоящей работе предлагается выбрать «круглое» значение $m = 0.5$ и ограничиться однократным оцениванием параметра Юла по модели (8).

### 4. Оценивание параметра распределения Юла без использования степенных распределений

Из выражений (1), (3) следует, что относительный прирост кумулятивной функции вероятности, равный отношению обычной функции вероятности к кумулятивной имеет вид

$$\delta_k = \frac{G_k - G_{k+1}}{G_{k+1}} = \frac{p_k}{G_{k+1}} = \frac{\theta - 1}{k}. \qquad (9)$$

Если выборка не содержит пропусков, т.е. массив значений $k_i$ имеет единичный шаг, то оценку параметра распределения Юла легко получить из (9) как сумму произведений $k_i$ на $\delta_i$, либо регрессией $\delta_i$ на $1/k_i$.

Отсутствие пропусков означает, что в выборке обязаны присутствовать макрообъекты со всеми возможными численностями составляющих их микрообъектов из заданного диапазона от $k_{min}$ до $k_{max}$. Например, если существуют статьи, процитированные сто раз, то должна быть хотя бы одна статья, процитированная 99 раз. Очевидно, что в реальных выборках это условие не выполняется. Например, для выборки, сгенерированной на рис.1, $k_i = [1...14, 16, 19, 20, 22, 24, 33, 48]$.

Выражение (9) можно обобщить на случай, когда шаг $d_i = k_{i+1} - k_i$ больше единицы:

$$\delta_i = \frac{p_{k_i}}{G_{k_{i+1}}} = \frac{\Gamma(k_i)\Gamma(k_i + \theta - 1 + d_i)}{\Gamma(k_i + d_i)\Gamma(k_i + \theta - 1)} - 1 = \prod_{j=k_i}^{k_{i+1}-1} \left(1 + \frac{\theta - 1}{j}\right) - 1 \qquad (10)$$

Учитывая, что в реальных выборках неединичные шаги $d_i$ находятся в диапазоне больших значений $k_i$ (численностей микрообъектов в макрообъектах), отбросим слагаемые второго и выше порядков (по $j$) в (10). Получим приближённое выражение

$$\delta_i \approx (\theta - 1) \cdot h_i, \qquad h_i = \sum_{j=k_i}^{k_{i+1}-1} \left(\frac{1}{j}\right). \qquad (11)$$

Для получения искомой оценки параметра распределения Юла построим линейную регрессию $\delta_i$ на $h_i$:

$$\hat{\theta}^{(4)} = 1 + \frac{\sum \delta_i h_i}{\sum h_i^2}. \qquad (12)$$

Следует отметить, что полученная оценка основана только на самом законе Юла (1), (3). Приближение этого распределения степенным не использовалось.

### 5. Сравнение точности полученных оценок параметра распределения Юла с помощью численного моделирования

Для сравнения качества оценок параметра распределения Юла, полученных разными методами, было проведено имитационное моделирование. Генерировалась сеть Барабаши-Альберт из $n = 1000$ вершин при истинном значении параметра $\theta = 3$, после чего строились оценки этого параметра: методом приближения к степенному распределению (традиционная log-log оценка $\theta^{(1)}$), скорректированная оценка $\theta^{(2)}$, оценки $\theta^{(3)}$, полученные на основе сдвинутого степенного распределения (при $m = 0.5$ и при «точном» значении $m(\theta = 3) = 0.7369$), и оценка $\theta^{(4)}$ (12), построенная на основе ряда отношений обычной функции вероятности к кумулятивной.

Полученные значения оценок были усреднены по $N = 100$ реализациям сети. В табл. 1

представлены средние значения и среднеквадратичные отклонения полученных оценок (относительно средних – $s(\theta)$ и относительно истинного значения – $\sigma(\theta)$ ).

**Таблица 1**

**Результаты оценивания параметра распределения Юла по данным численного эксперимента**

| | $\theta^{(1)}$ | $\theta^{(2)}$ | $\theta^{(3)}_{m=1/2}$ | $\theta^{(3)}_{m=m_3}$ | $\theta^{(4)}$ |
|---|---|---|---|---|---|
| Среднее | 2,7124 | 2,8332 | 2,9194 | 2,9200 | 2,9833 |
| $s(\theta)$ | 0,0739 | 0,0739 | 0,1021 | 0,1020 | 0,1198 |
| $\sigma(\theta)$ | 0,2983 | 0,1832 | 0,1304 | 0,1299 | 0,1210 |

Данные результаты показывают, что смещение оценки $\theta^{(4)}$ (12) является минимальным среди всех приведенных. Кроме того, можно утверждать, что оценка $\theta^{(3)}$, построенная на основе сдвинутого степенного распределения, предпочтительнее оценок $\theta^{(1)}$ и $\theta^{(2)}$. Причём величина сдвига ( $m$ ) слабо влияет на точность этой оценки.

## Выводы

В настоящей работе проведён анализ области применения модели Юла и методов оценивания параметра распределения Юла-Саймона. Он показал, что традиционный метод log-log регрессии, основанный на приближении закона Юла степенным, порождает смещённые оценки, а метод, предложенный Гарсиа [6], характеризуется высокой вычислительной сложностью.

Предложены три оригинальных метода оценивания искомого параметра: коррекция log-log оценки (аддитивным фактором, зависящим от объёма выборки), оценивание параметра Юла с помощью сдвинутого степенного распределения и оценивание на основе ряда отношений обычной функции вероятности к кумулятивной.

Проведён численный эксперимент, в котором полученные методы оценивания сравнивались по точности на модельных объектах – сетях Барабаши-Альберт. Результаты моделирования наглядно продемонстрировали преимущество последних из упомянутых оценок – единственных, при построении которых не использовалось приближение закона распределения Юла дискретным степенным.

Очевидным направлением дальнейших исследований должно стать доказательство несмещённости оценки, построенной на основе ряда отношений обычной функции вероятности к кумулятивной.

**Список литературы:**

**1.** *Dorogovtsev, S.N.* Evolution of Networks: From Biological Networks to the Internet and WWW [Text] / Dorogovtsev S.N., Mendes J.F.F. – Oxford, USA: Oxford University Press, 2003. – 280 p. **2.** Ландэ, Д.В. Моделирование сложных сетей: учебное пособие [Текст] / А.А. Снарский, Д.В. Ландэ — К. : НТУУ «КПИ», 2015. — 212 с. **3.** *Albert, R.* Statistical mechanics of complex networks [Text] / Albert R., Barabasi A.- L : Rev. Mod. Phys. - 2002. - V. 74. - p. 42-97. **4.** *Newman, M.E.J.* Power laws, Pareto distributions and Zipf's law [Text] / Mark Newman. – Contemporary Physics, 2005, 46(5). p.323-351. doi: 10.1080/00107510500052444. **5.** *Олемской А.И.* Статистика сложных сетей (обзор) [Text] / А.И. Олемской, И.А. Олемской - «Вісник СумДУ». – 2006. – №6 (90). – С.21-47. **6.** *Garcia G.J.M.* A fixed-point algorithm to estimate the Yule-Simon distribution parameter [Text] / Garcia Garcia J.M. – Applied Mathematics and Computation, 2011, 217, p.8560-8566. **7.** Масштабно-инвариантные сети [Электронный ресурс] – Режим доступа: http://www.cognitivist.ru/er/kernel/power_laws_2.xml – 2014 г. – Загл. с экрана.

***Поступила в редколлегию 18.05.2017***