Analysis and Simulation of Power Law Distribution of File Types in File Sharing Systems

Yuya Dan
Faculty of Business Administration
Matsuyama University
Bunkyo 4-2, Matsuyama, Ehime 790-8578, Japan
Email: dan@cc.matsuyama-u.ac.jp

Takehiro Moriya
CTO and R&D Headquarters
Branddialog, Inc.
Minato 3-5-10, Chuo-ku, Tokyo 104-0043, Japan
Email: moriya@branddialog.co.jp

***Abstract*—We study the distribution of file types classified by file extensions in usual file systems. In this paper, we report that the power-law distribution is observed in a certain file system and try to give the answer to the mechanism of that formation. In order to recognize the phenomena, we construct mathematical models and compare them to the results of Monte Carlo simulation. Then, we propose that file operation of creation and copy would form the distribution at the conclusion. This paper focuses on the formation of the power-law distribution by mathematical analysis and computer simulation.**

***Keywords*-power-law distribution; mathematical modeling; Monte Carlo simulation; file operation; scale-free structure**

I. INTRODUCTION

In a variety of sciences, we observe that the value distributes a typical value around the average which individual measurements are centered. Gaussian distributions are often obtained when scientists measure their targets. Despite of Gaussian distributions, there are binomial, Poisson and power-law distributions in observed scientific data. In particular, power-law distributions in the observation has no typical value as averages, so that we also call scale-free structure.

There are many examples of distributions that obey power-law in natural, social, and other sciences. We know Gutenberg-Richter's law as the sizes of earthquakes [13], Zipf's law as the frequency of use of words in any human language [27], the numbers of papers scientists write [15], the number of citations received by papers [20], the number of hits on web pages [1], structure of WWW traffic [7], people's annual incomes [19], the sales of music recordings [5], the frequency of opening moves in chess [4], and so on. See also city populations and the property of power-law phenomena [16] more in detail. Clauset et. al. [6] gave a concise statistical method for analysis of power-law phenomena.

Mathematicians and physicists would believe that comparative simple principles form complex structure in these fields, and try to recognize the essential framework of models they proposed. In fact, We know that chaotic phenomena often occur even in the simple system. It is natural that complex systems are made from the simple principles.

In this paper, we propose a model for file operation process as a complex system, then provide a simulation result based on the model. File operation process is one of human-computer interaction in our ordinary computer life, that we unwittingly create, copy, move and deleted the files in our storages. It seems to be a random process that we do such file operations, however, we can obtain the highlight data which occur in the file operation process.

This paper is organized into five sections. Section II gives a brief review for the result of observation in a certain file sharing system. After motivated, Section III describes the construction of the mathematical model for file operations. The simulation of the proposed model is presented in section IV. Finally, Section V concludes the paper.

II. OBSERVATION

We found out the distribution of file types in the file sharing system of social groupware "GRIDY" [12] that is used by over 10,000 registered companies in Japan. They share files on the cloud, that is a virtual storage on the Internet. File types can be classified by their file extensions, so that we have statistics of file types. Figure 1 shows the doubly logarithmic plot of the frequency of each file type by descending order. $p(k)$ is defined the number of files in the same extension at the k-th order. It is easy to see from regression that the distribution of file types seem to follow power-law distribution;

$$p(k) = Ck^{-\gamma} \tag{1}$$

with $C = 780,359$ and $\gamma = 2.438$ at the comparative high coefficient of determination in $R^2 = 0.9864$. This is our motivation of discussion why power-law distribution forms in file types.

In more detail, the data fit near the regression curve (1) although the data at small k are far from the law of power. There are 264 file types and the largest number of files at $k = 1$ is 63,392.

 ISBN: 978-1-61208-169-4

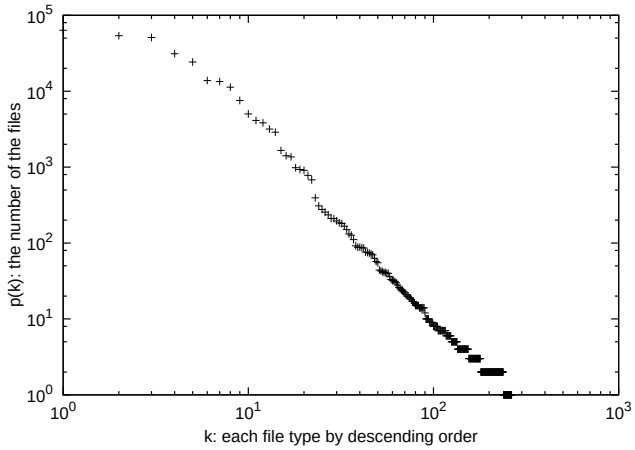

Figure 1. Observed distribution of file types in a file sharing system

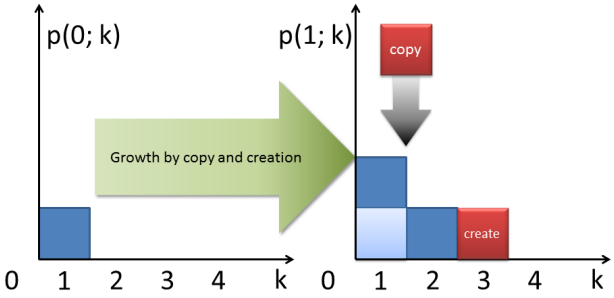

Figure 2. A model of file operation

III. Mathematical Model for File Operation

In this section, we construct a mathematical model for file operation, and show a fine result of power-law distribution of file types.

First of all, we define $p(t;k)$ as an integer-valued discrete function of t and k. The variable t means the time step which may take $\{0,1,2,\ldots\}$ and each k represents a sort of file types which may take $\{1,2,\ldots\}$. There is one file in the system when $t=0$. At the next step, we copy a file from the existing file to the system which type is $k=1$ and we create a file at $k=2$. When we copy a file from the existing files, we select a type of files at the probability proportional to the number $p(t;k)$ of the existing files in the system. Adding to this, we create a file at $p(t;k)$. See also Figure 2 in detail.

Mathematical analysis for the model indicates

$$p(t+1;k) = p(t;k) + \frac{p(t;k)}{2t+1}\cdot 1, \tag{2}$$

Table I
Estimation of Stirling's approximation

n	$n!$	Stirling	ratio
1	1	0.92	0.922
2	2	1.92	0.960
5	120	118	0.983
10	3628800	3.60×10^{6}	0.992
20	2.43×10^{18}	2.42×10^{18}	0.996
50	3.04×10^{64}	3.04×10^{64}	0.998
100	9.33×10^{157}	9.32×10^{157}	0.999

where we used

$$\sum_k p(t;k) = 2t+1 \tag{3}$$

as the sum of all possible k. The boundary condition for file creation, we can write

$$p(t;t+1) = 1 \tag{4}$$

for all t and

$$p(t;k) = 0 \tag{5}$$

for every k with $k \geq t+2$. Since the recurrence relation

$$\frac{p(t;k)}{p(t-1;k)} = \frac{2t}{2t-1} \tag{6}$$

and initial value

$$p(k-1;k) = 1, \tag{7}$$

we obtain

$$\begin{aligned} p(t;k) &= \frac{2t}{2t-1}\cdot\frac{2t-2}{2t-3}\cdot\cdots\cdot\frac{2k}{2k-1} \\ &= \prod_{j=k}^{t}\frac{2j}{2j-1} \\ &= \prod_{j=1}^{t}\frac{2j}{2j-1}\Bigg/\prod_{j=1}^{k-1}\frac{2j}{2j-1}. \end{aligned} \tag{8}$$

In our discussion, we use Stirling's approximation

$$n! \sim \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n} \tag{9}$$

for sufficiently large n. See also Table I for accuracy of Stirling's approximation.

Applying Stirling's approximation to the previous expression, we have

$$\begin{aligned} \prod_{j=1}^{t}\frac{2j}{2j-1} &= \frac{2^t t!}{\frac{(2t)!}{2^t t!}} \\ &= \frac{2^{2t}(t!)^2}{(2t)!} \\ &\sim \frac{2^{2t}(2\pi)t^{2t+1}e^{-2t}}{\sqrt{2\pi}(2t)^{2t+\frac{1}{2}}e^{-2t}} \\ &= \sqrt{2\pi}t^{\frac{1}{2}}, \end{aligned} \tag{10}$$

so that it is concluded that the limit of the expression around t converges to $\sqrt{2\pi}$;

$$\lim_{t\to\infty} t^{-\frac{1}{2}} \prod_{j=1}^{t} \frac{2j}{2j-1} = \sqrt{2\pi}. \tag{11}$$

Similarly, we have

$$\begin{aligned} \prod_{j=1}^{k-1} \frac{2j}{2j-1} &= \frac{2^{k-1}(k-1)!}{\frac{(2k-2)!}{2^{k-1}(k-1)!}} \\ &= \frac{2^{2k-2}((k-1)!)^2}{(2k-2)!} \\ &\sim \sqrt{\pi}(k-1)^{\frac{1}{2}} \end{aligned} \tag{12}$$

for sufficient large k.

Summing up these calculations, we conclude

$$\lim_{t\to\infty} t^{-\frac{1}{2}} \prod_{j=1}^{t} \frac{2j}{2j-1} \Bigg/ \prod_{j=1}^{k-1} \frac{2j}{2j-1} = \sqrt{2}(k-1)^{-\frac{1}{2}}, \tag{13}$$

which indicates the power-law distribution.

IV. Computational Experiment

In order to investigate power-law distribution, we construct a network on the computer.

A. Simulation

There is a vertex at the beginning of the simulation. We construct the network by adding vertices according to the probabilities proportional to the number of edges that the candidate vertex have. In other words, a person who have many friends is tend to have new friends. In the simulation, we can see evolution of networks that have at most $8,048$ vertices.

Figure 3 summarized the flow chart of procedure in our simulation. First of all, every element of the array are initialized, and put the first vertex on the network. Next, the program loops it until the number of vertices is 8,084 that a new vertex comes to the network and select a vertex to be connected by the application of preference selection. The number $8,084$ can be extended to $8,084^2$. We restricted the maximum number of the array for the reason of analysis the network structure, however, the restriction is not necessary for the case that we see the number of edges in the network. At the last stage, the program make the histogram for the number of edges, then we can estimate the distribution by regression in statistical analysis.

The source code of the simulation is written in Java which is shown in the appendix at the last part of the paper, and the program ran on Intel Core 2 Duo CPU (T9600 @2.80GHz x2) with Microsoft Windows Vista 32bit version.

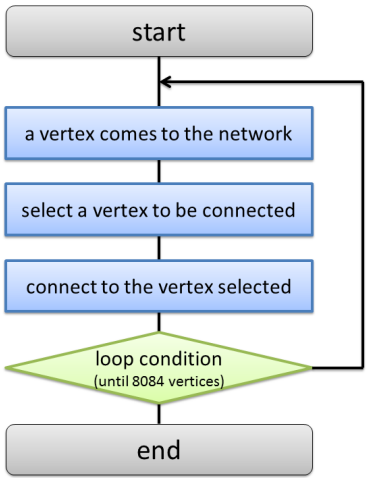

Figure 3. Flow chart of the simulation

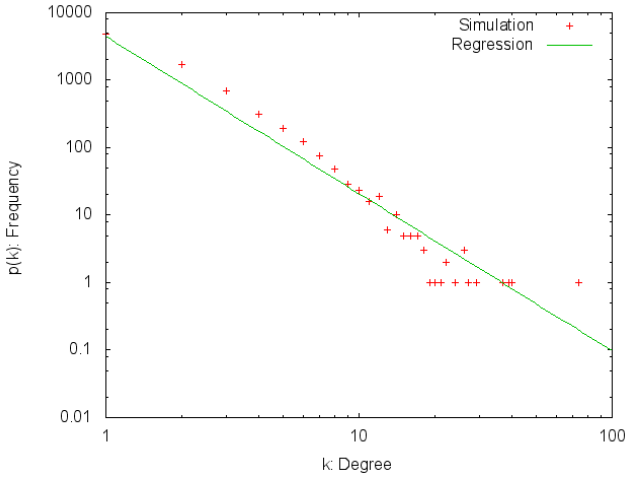

Figure 4. Distribution of degrees of vertices and regression curve

B. Results

Figure 4 shows the distribution of degrees of each vertex with regression curve. We use logarithmic scale both in axis. The result indicates

$$p(k) = 4.47 \times 10^3 k^{-2.33}, \tag{14}$$

which is characterized by scale-free structure of networks with $\gamma = 2.33$. The result of a trial is outputted as follows:

```
# === Simulation ===
# a = -2.3317937664992416
# b = 8.40563662841364
# R^2 = 0.8875206382378562
1 4757
2 1702
3 701
```

V. CONCLUSION

According to the relevant results [16], we expect the power-law distribution of $\gamma = 3$ by the application of preference selection. In our result, we obtain $\gamma = 0.5$ from mathematical estimate, and $\gamma = 2.33$ from computer simulation. There is still a gap between mathematical analysis and simulation results. In addition to this result, we have the property of γ to converse to 3 if we give a large number of vertices to the network.

We have studied construction of scale-free networks in stochastic process. According to the model proposed by Barabási and Albert, we can construct scale-free networks using connecting probability that is proportional to the number of edges each vertex has.

In the problem of file types, we assume to copy files from old ones at random, so that we can conclude there is a similar effect to construct scale-free networks and power-law distribution emerges in file operations.

ACKNOWLEDGMENT

The authors would like to thank the referees for their useful comments on the previous manuscript of the paper.

REFERENCES

[1] L. A. Adamic and B. A. Huberman, "The Nature of Markets in the World Wide Web," *Q. J. Electron. Commerce*, Vol. 1, pp. 5–12. (2000)

[2] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. Mod. Phys.*, Vol. 74, No. 1, pp. 47–97. (2002)

[3] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science* 286, pp. 509–512. (1999)

[4] B. Blasius and R. Tönjes, "Zipf's Law in the Popularity Distribution of Chess Openings," *Phys. Rev. Lett*. Vol. 103, pp. 218701. (2009).

[5] R. A. K. Cox, J. M. Felton, and K. H. Chung, "The Concentration of Commercial Success in Popular Music: An Analysis of the Distribution of Gold Records," *Journal of Cultural Economics*, Vol. 19, pp. 333–340. (1995)

[6] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review* 51 (4), pp. 661–703. (2009)

[7] M. E. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes," in *Proceedings of the 1996 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pp. 148–159. (1996)

[8] Y. Dan, "Modeling and Simulation of Diffusion Phenomena on Social Networks," *The Proceedings of 2011 Third International Conference on Computer Modeling and Simulation (ICCMS 2011)*, pp. 139–146. (2011)

[9] Y. Dan, "Mathematical Analysis and Simulation of Information Diffusion on Networks," *The Proceedings of The 11th IEEE/IPSJ International Symposium on Applications and the Internet (SAINT 2011)*, pp. 550–555. (2011)

[10] S. N. Dorogovtsev, *Lectures on Complex Networks*, Oxford University Press. (2010)

[11] Dorogovtsev, S. N., J. F. F. Mendes, and A. N. Samukhin, "Structure of Growing Networks with Preferential Linking," *Phys. Rev. Lett.* 85, pp. 4633–4636. (2000)

[12] GRIDY, http://gridy.jp/

[13] B. Gutenberg and R. F. Richter, "Frequency of Earthquakes in California," *Bull. Seismol. Soc. Am. 34 185*, Vol. 34, no. 4, pp. 185–188. (1944)

[14] L. Kullmann and J. Kertész, "Preferencial Growth: Exact solution of the time dependent distributions," *Phys. Rev. E 63*, 051112. (2001)

[15] A. J. Lotka, "The Frequency Distribution of Scientific Productivity," *J. Wash. Acad. Sci.* Vol. 16, pp. 317–323. (1926)

[16] M. E. J. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, Vol. 46, no. 5, pp. 323–351. (2005)

[17] M. E. J. Newman, *Networks*, Oxford University Press. (2010)

[18] M. E. J. Newman, A. L. Barabási, and D. J. Watts, *The Structure and Dynamics of Networks*, Princeton University Press. (2006)

[19] V. Pareto, *Cours d'Économie Politique*, Droz, Geneva. (1896)

[20] D. J. de S. Price, "Networks of Scientific Papers," *Science*, 149 pp. 510–515. (1965)

[21] W. Reed and B. D. Hughes, "From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature," *Physical Review E*, Vol. 66, Issue 6, id. 067103. (2002)

[22] E. M. Rogers, *Diffusion of Innovations, 5th ed.*, Free Press, New York. (2003)

[23] H. A. Simon, "On a class of skew distribution function," *Biometrika*, Vol. 42, pp. 425–440. (1955)

[24] A. Vázquez, R. Pastor-Satorras, and A. Vespignani, "Large-scale topological and dynamical propertes of the Internet," *Physical Review E*, Vol. 65, No. 066130. (2002)

[25] P.-F. Verhulst, "Notice sur la loi que la population poursuit dans son accroissement," *Correspondance mathématique et physique* 10, pp. 113–121. (1838)

[26] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature* 393, pp. 440–442. (1998)

[27] G. K. Zipf, *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, Cambridge. (1949)

 ISBN: 978-1-61208-169-4

APPENDIX

Here is the Java source code used for the numerical simulation in our research:

```
/*
  Simulation for Emergence in Complex Networks
  Copyright (C) 2011 Yuya Dan, Matsuyama University */

import java.util.Random;

public class NetworkEmergence{
  public static void main( String[] args ){
    final int NUM = 8048;  // Max{NUM} on memory = 8048
    Random r = new Random( 0 );

    // Start the Simulation
    System.out.println( "# === Emergence Simulation in Networks ===" );

    // Initialization
    int n = 1;
    int[] a = new int[NUM];

    for( int i = 0; i < a.length; i++ ){
      a[i] = 1;
    }

    // Construction of a Network
    for( int i = 0; i < NUM - 1; i++ ){
      int sum = 0;
      for( int j = 0; j < n; j++ ){
        sum += a[j];
      }
      int x = r.nextInt( sum );
      int s = 0;
      int j;
      for( j = 0; j < n && s <= x; j++ ){
        s += a[j];
      }
      a[--j]++;
      a[n]++;
      n++;
    }

    // Make the Histogram
    int max = 0;
    for( int i = 0; i < a.length; i++ ){
      if( max < a[i] ){
        max = a[i];
      }
    }
    int[] histogram = new int[max];
    for( int i = 0; i < histogram.length; i++ ){
      histogram[i] = 0;
    }
    for( int i = 0; i < a.length; i++ ){
```

```
      histogram[a[i] - 1]++;
    }

    // Statistical Analysis
    int nn = 0;
    double ax = 0.0, ay = 0.0;
    for( int i = 0; i < histogram.length; i++ ){
      if( histogram[i] > 0 ){
        ax += Math.log( (double)i );
        ay += Math.log( (double)histogram[i] );
        nn++;
      }
    }
    ax /= nn;
    ay /= nn;

    double Sxx = 0.0, Syy = 0.0, Sxy = 0.0;
    nn = 0;
    for( int i = 0; i < histogram.length; i++ ){
      if( histogram[i] > 0 ){
        Sxx += ( Math.log( (double)i ) - ax )
               * ( Math.log( (double)i ) - ax );
        Sxy += ( Math.log( (double)histogram[i] ) - ay )
               * ( Math.log( (double)i ) - ax );
        Syy += ( Math.log( (double)histogram[i] ) - ay )
               * ( Math.log( (double)histogram[i] ) - ay );
        nn++;
      }
    }
    Sxx /= nn;
    Sxy /= nn;
    Syy /= nn;
    System.out.println( "# a = " + ( Sxy / Sxx ) );
    System.out.println( "# b = " + ( ay - Sxy / Sxx * ax ) );
    System.out.println( "# R^2 = " + ( Sxy / Sxx ) * ( Sxy  / Syy ) );

    // Result
    for( int i = 1; i < histogram.length; i++ ){
      System.out.println( i + "\t" + histogram[i] );
    }
  }
}
```

 ISBN: 978-1-61208-169-4