送信された電子メールのサイズ頻度のメールヘッダの内容に基づく分析

松原 義継[1,2,a)]　武藏 泰雄[3,b)]

概要：電子メールサイズの頻度分布を説明するためのモデルを構築した．とある大学の電子メール送信サーバへリクエストされた電子メールのサイズの頻度分布図を 100 バイト単位で作成した．その頻度分布の大きな形からは，べき則を読み取れる．電子メールヘッダの 1 つである “Content-Type” の内容に基づき，その頻度分布を 4 種類に分解した．この分解に用いた電子メールの送信リクエスト回数は 269,085 (3 ヶ月分) である．分解された各頻度分布を説明するためのモデルを作成した．提案モデルでは，送信者は意識的もしくは無意識的に電子メール本文中の新規作成文の桁を正規分布に従い管理する．そのモデルからは概ね受け入れられるフィッティング結果を得られ，さらに実験によりべき則性を有することの結果も得られた．

キーワード：電子メールサイズ，べき則分布，メールヘッダ

Analysis of frequency distribution in email size based on an email header

Matsubara Yoshitsugu[1,2,a)]　Musashi Yasuo[3,b)]

Abstract: We propose a model to explain frequency distribution in email size. We made frequency distribution of email size in the system log of the staff email server at a university campus. Power-law properties are recognized in the distribution. We disaggregated the distribution to four subdistributions based on the “Content-Type” email header. The total number of disaggregated emails is 269,085. Then, we propose a model to explain each subdistribution that obeys a log-normal-like distribution. In this model, email senders – consciously or unconsciously – manage the size of new sentences, obeying a normal distribution. The fit of our model is acceptable, and the model demonstrates power-law properties for large email sizes.

Keywords: E-mail sizes, Power-law distribution, Gamma distribution

1. はじめに

インターネットは，現代社会において広く用いられている通信媒体の 1 つである．その構造や流量の動的性質を研究対象とする動きがある．例えば，AS レベルでのインターネットの構造にはスケールフリー性 [1]，パケット流量の時系列上にはべき則に従う相関 [2–11]，イベント間の時間間隔にはべき則 [6–11] が見出されている．

これら報告の中に，電子メール送信サーバにリクエストされた電子メールのサイズの頻度分布にべき則を見出した報告がある [12]．その報告では電子メールサイズの頻度分布がべき則に従う場合，電子メールのデータ流量の時系列はランダム化することが実験により示されている．文献 [13] は，そのような頻度分布を説明するモデルの提案が

[1] 佐賀大学
Saga University, 1 Honjo-machi, Saga-shi, Saga, 840-8502, Japan
[2] 熊本大学 大学院
Graduate School of Science and Technology, Kumamoto University, 2-40-1 Kurokami Chuo-ku, Kumamoto-shi, Kumamoto, 860-8555 Japan
[3] 熊本大学
Kumamoto University, 2-40-1 Kurokami Chuo-ku, Kumamoto-shi, Kumamoto, 860-8555 Japan
[a)] matubara@cc.saga-u.ac.jp
[b)] musashi@cc.kumamoto-u.ac.jp

行われている．

今回，我々は文献 [13] での分析を更に推し進めた．具体的には，頻度分布における電子メールのサイズの単位の変更および頻度分布の構成要素への分解である．これらを基に，その分解された各頻度分布を説明するためのモデルを提案する．

始めに，電子メールのサイズの小さい領域での構造を見出すために，我々は頻度分布における電子メールのサイズの区間分割単位をこれまでの 1 キロバイトごとから 100 バイトごとに変更する．1 キロバイト未満のサイズの電子メールは一定数以上存在することから，それらの領域の構造も考慮した分析を行う．

この変更された単位を基に，頻度分布の形を再分析する．再分析に用いる電子メールサーバは文献 [12] で用いたのと同一であり，電子メールの送信リクエスト回数は 5,536,003 (6 年間分) である．

さらに，電子メールのヘッダの 1 つである “Content-Type” の内容に基づき，頻度分布の分解を行う．我々は，観測されている頻度分布に現れている変曲点に注目し，頻度分布の形は異なる種類の電子メールサイズの頻度分布の合成であるという仮設を設ける．今回の分析では，電子メールの種類は “Content-Type” の内容に基づき 4 種類に設定する．電子メールサーバのログファイルには，通常は電子メールのヘッダ部は記録されない．そこで，データを収集する電子メールサーバを運用している組織から “Content-Type” ヘッダの収集許可を頂き，電子メールサーバに送信された各電子メールの “Content-Type” ヘッダの内容を収集した．収集できた電子メールの内で各種類へ分類可能だった電子メールリクエスト回数は 269,085 (3 ヶ月分) である．

これら分析結果を踏まえ，各種類へ分解した各頻度分布を説明するためのモデルの提案および考察を行う．

利用者の様々な理由により送信された電子メールの全体としての性質を分析することにより，電子メール送信というネットワークサービスの利用に関する統計的知見を得ることができる．それら知見の中には，サービスの実運用において有益な知見が含まれていることが期待される．

2. 分析

始めに，2009 年度から 2014 年度までの各年度において，とある大学の教職員用電子メールサーバへリクエストされた各電子メールのサイズの頻度分布を年度 (4 月 1 日から翌年 3 月 31 日) 単位で集計する．集計単位を年単位ではなく年度単位にする理由は，電子メールの利用者側の環境は年度単位で大きな変化を生じると考えられるからである．毎年 4 月になると，教職員は人事移動が多く行われる．退職もしくは転勤等により教職員の一部は変更となり，さらに別の部局への内部異動や昇進等により教職員の一部は業務内容の変更を生じさせる．これらのことから，人事異動は電子メールの利用者側の環境に大きな変化をもたらすと考えられる．この事を鑑みて，本論文では年度単位で集計する．

電子メールのサイズの頻度を集計する際，そのサイズの単位は先行研究 [12,13] での 1 キロバイト [kB] から 100 バイトへ詳細化する．理由は，1kB 未満のサイズの電子メールは一定数以上存在しており，その領域の構造を分析するには 1kB 単位では頻度分布形が粗いからである．しかしながら，1 バイト単位では頻度分布形は不鮮明になり分析困難であった．従って，今回の分析では電子メールのサイズの単位は 100 バイトに設定する．作成した頻度分布を **図 1** に示す．図内での表記 ‘AY’ は 年度 (academic year) の略記である．各図の横軸は電子メールのサイズ (100 バイト単位)，縦軸は各サイズに対応した電子メールの頻度である．両軸は対数化されている．図中の各青点は実際の値を表す．図中の各頻度分布からは電子メールサイズ 1 kB ($s = 10$) 付近でピークを読み取れ，15 kB および 40 kB 付近に変曲点を読み取れる．40 kB よりサイズの大きい領域では各頻度分布の形から直線性を読み取れる．各頻度分布は両対数グラフであることから，各頻度分布で直線性を読み取れる領域はべき則に従っていることを意味している．

2.1 頻度分布の各構成要素への分解

図 1 の各頻度分布から読み取れる電子メールサイズ 15 kB および 40 kB 付近での変曲点の原因について，我々は種類の異なる電子メールの頻度分布の合成結果と考えている．現実の電子メール作成の場合，文章のみで数十キロバイトを超え数メガバイトに至る電子メールを作成することは困難である．そこで，この考えを基に電子メールサイズの頻度分布を電子メールの種類毎に分解する．

電子メールの種類を決定するために我々は RFC にて定義されている電子メールに関する各種プロトコルの 1 つである MIME (Multipurpose Internet Mail Extension) [14–19] に注目した．MIME にて定められている各規約の中に電子メールの中身の種類を定めた “Content-Type” が存在する．例えば，平文メールはヘッダ部に “Content-Type: text/plain;” を含む行が存在する．もし電子メールに何かしらのファイルが添付されている場合は，そのヘッダ部には “Content-Type: multipart/” を含む行が存在し，添付ファイルの種類を表す “Content-Type: ” で始まる行は文章の含まれているボディ部に記述される．我々は，今回の分解のために この “Content-Type” を利用する．

電子メールの種類の決定方法は複数考えられるが，本研究において各電子メールは **表 1** に基づき 4 種類に分解する．具体的には平文 (“Plain”)，HTML 形式 (“HTML”)，テキストファイルの添付付き (“Text attachment”)，そしてバイナリファイルの添付付き (“Binary attachment”) で

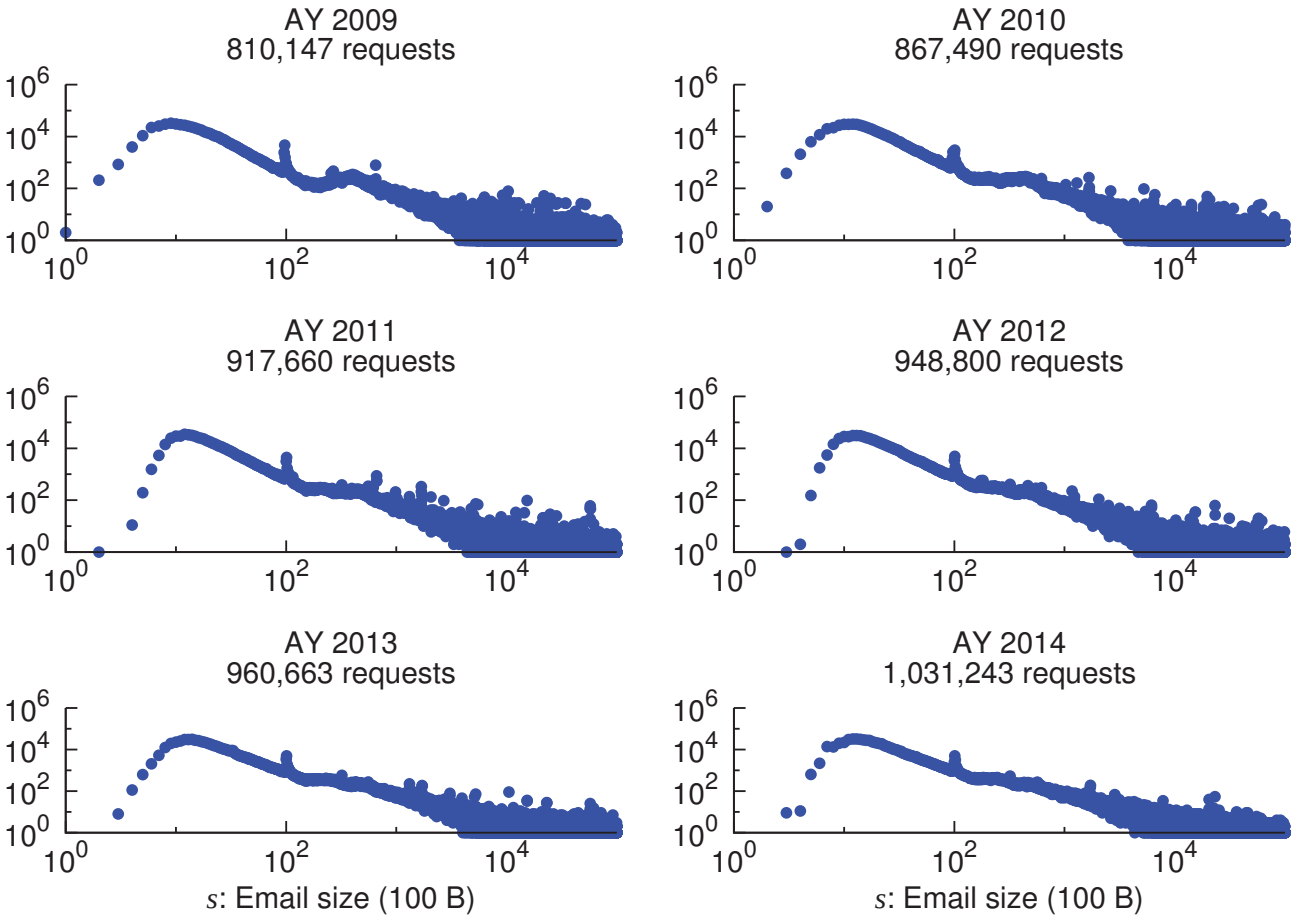

図 1　教職員用電子メール送信サーバにおける電子メールサイズ頻度の年度別変化.
Fig. 1　Frequency in e-mail sizes for staff per academic year.

ある．いくつかの電子メールは “Content-Type” ヘッダを有さないもしくは “Content-Type” の具体の記述が規約に反しており，それらは分析対象外とした．

リクエストされた各電子メールの “Content-Type” 情報は，通常は電子メールサーバのログファイルには記録されない．そこで，我々はこの “Content-Type” 情報を収集する許可を得て，電子メールサーバのログファイルに その情報を記録されるように電子メールサーバに設定を施した．“Content-Type” 情報を含む電子サーバへのリクエスト記録は，2015 年 5 月 1 日から同年 7 月 31 日までの期間に教職員用電子メール送信サーバで収集された．その期間に収集された電子メールのサイズ頻度分布を図 **2** に示す．縦軸は電子メールサイズ s の値に対応する頻度である．この期間における電子メールのリクエスト回数は 272,986 であった．“Content-Type” 情報の欠落もしくは規約に反した記述により，3,901 回のリクエストは分析対象外となった．従って，分析対象となった電子メールリクエスト回数は 269,085 である．図 2 からも 2 つの変曲点を読み取れ，s の値の大きな領域ではべき則性も読み取れる．

構成要素に分解された各頻度分布を図 **3** に示す．縦軸は電子メールサイズ s の値に対応する頻度である．各頻度分布は色分けされている．各頻度分布における電子メールリクエスト回数は，それぞれ “Plain” は 158,913，“HTML” は 33,080，“Text attachment” は 4,943，“Binary attachment” は 72,149 である．図 2 および図 3 を併せて鑑みると，観測データ全体の頻度分布の形への影響が強いのは “Plain” および “Binary attachment” と考えられる．更なる議論は 4 節にて行う．

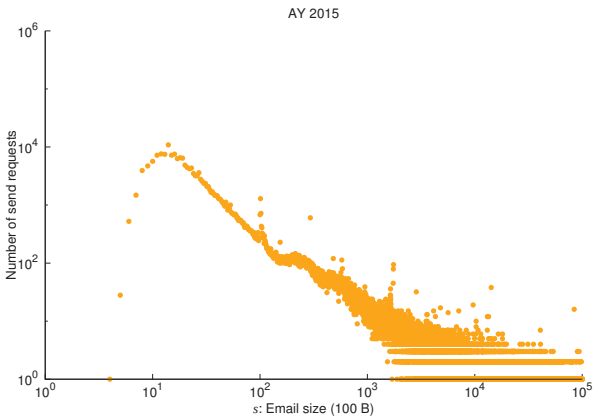

図 2　電子メールサイズの頻度分布.
Fig. 2　Frequency distribution of Email sizes.

表 1　“Content-Type” に基づく電子メールの分類.
Table 1　Types of emails based on the “Content-Type” email header.

Group	Email type	Content	Construction of “Content-Type” headers based on MIME protocol
No attachment	Plain	Plain text	One “text/plain” only.
	HTML	HTML formatted	The first is “multipart/alternative,” the second is “text/plain,” the last is “text/html.”
Attachment	Text attachment	Attached text files	The first includes a primary type “multipart/,” the others include the primary types “text/” and/or “message/.”
	Binary attachment	Attached binary files	The first includes a primary type “multipart/,” the others include the primary types “application/,” “image/,” “audio/,” and/or “video/.”

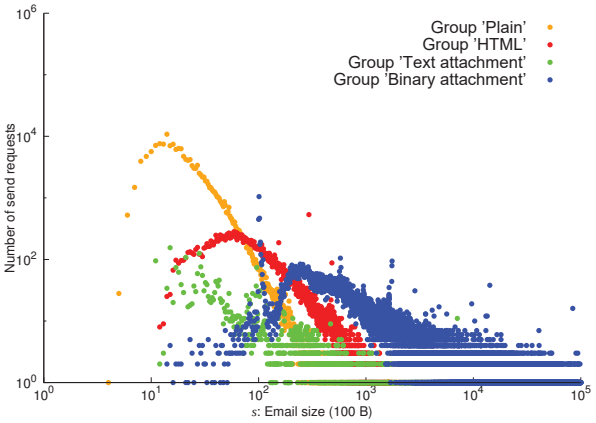

図 3　各構成要素へ分解された頻度分布.
Fig. 3　The disaggregated frequency subdistribution.

3. モデル式の提案

これまでの分析を踏まえて図 3 の各頻度分布形を説明可能なモデル式を以下に提案する.

$$p(s) = \frac{1}{a}\frac{1}{s\ln s}\exp\left\{\frac{-(\ln\ln s-\mu)^2}{2\sigma^2}\right\}, \qquad (1)$$

ここで a (>0) は正規化定数，$-\infty<\mu<\infty$, $\sigma>0$ である．理論上，s の値の上限は無制限である．そこで，$p(s)$ を s についての連続系とみなして $a=\sqrt{2\pi}\sigma$ とする．

式 1 は，対数正規分布 $f(x)$ および 式 $x=\ln s$ の合成である．$f(x)$ は電子メールの本文中における新規作成文の長さ x の頻度分布とする．x の対数形 $\ln x$ は x の桁 (スケール) に相当し，その頻度は正規分布に従うとする．これは，$f(x)$ が対数正規分布であることに由来する．提案モデルでは，電子メール送信者は意識的もしくは無意識的に電子メールの桁を扱っていると考える．電子メールにおけるヘッダ部，返信元電子メールからの引用文，HTML 形式電子メールでの HTML タグ，そして添付ファイルは x の余剰部分として考える．そして，それら余剰部分のサイズは新規作成文のサイズよりも明らかに大きいとする．この考えを基に x および s との間の関係は $x=\ln s$ として表される．実際の電子メールにおける $x=\ln s$ の例として例えば以下の 2 つが考えられる．

(1) ある 2 ユーザ (A, B) が平文の電子メールをやり取りすることを考える．始めに，A は平文の電子メールを新規作成して B へ送信する．その電子メールを受け取った B は，A へ元の本文の全文もしくは一部を引用文として残して返信する．その返信電子メールを受け取った A は，B からの電子メールの本文の全文もしくは一部を引用文として残して B へ返信する．この返信を何回も繰り返していくことにより，引用文のサイズは新規作成文のサイズよりも明らかに大きくなる．

(2) あるユーザはデジタルカメラやスマートフォン等で撮影した写真を電子メールに添付して送信する．最近のこれらの機器で撮影した写真ファイルのサイズは，容易に数百キロバイトに達する．新規作成文のみで数百キロバイトに達するのは困難である．

提案モデルおよびべき則との関係について考察する．s の値が非常に大きい場合 $(s\gg 1)$ について，提案モデル式 1 の両辺を対数化すると次のように近似できる可能性がある．

$$\begin{aligned}\ln p(s) &\sim -\ln s-\frac{(\ln\ln s)^2}{2\sigma^2}\\ &\sim -\ln s.\end{aligned}$$

各 s の値に対応する $\ln s$ および $(\ln\ln s)^2$ の値の大小関係は $\ln s>(\ln\ln s)^2$ であることから，$s\gg 1$ の場合に $p(s)\sim s^{-1}$ でべき則に従うと考えられる．$p(s)$ のべき則性に関する更なる議論は 4 節で行う．

x の値が非常に大きい場合 $(x\gg 1)$ について，対数正規分布 $f(x)$ の両辺を対数化すると次のように近似できる．

$$\begin{aligned}\ln f(x) &\sim -\ln x-\frac{(\ln x)^2}{2\sigma^2}\\ &\sim -(\ln x)^2.\end{aligned}$$

この $f(x)$ の近似は $\ln x$ の 2 次式であることから，$f(x)$ はべき則性を有さないことを意味する．故に，$p(s)$ がべき則性を有するならば，それは $x=\ln s$ によりもたらされると考えられる．

構成要素に分解された各頻度分布の合成系である観測データのモデル式 $p_{\mathrm{S}}(s)$ は

$$p_{\mathrm{S}}(s) = p_{\mathrm{P}}(s)p(\mathrm{P}) + p_{\mathrm{H}}(s)p(\mathrm{H}) + p_{\mathrm{T}}(s)p(\mathrm{T}) + p_{\mathrm{B}}(s)p(\mathrm{B})$$

として表される．ここで，$p_{\mathrm{P}}(s)$，$p_{\mathrm{H}}(s)$，$p_{\mathrm{T}}(s)$，そして $p_{\mathrm{B}}(s)$ はそれぞれ “Plain”，“HTML”，“Text attachemnt”，そして “Binary Attachment” の頻度分布を表す．電子メールサーバへの送信リクエスト 1 回に付き，全送信に占める各要素の割合に応じて いずれか 1 種類がランダムに選択される．$p(\mathrm{P})$，$p(\mathrm{H})$，$p(\mathrm{T})$，そして $p(\mathrm{B})$ は その事を意味している．そして，選択された頻度分布でサイズが確率的に決定される．

4. 議論

提案モデル式 1 により生成される電子メールサイズの頻度分布形における実際の頻度分布形とのフィッティングの程度およびべき則性を議論する．

4.1 頻度分布とのフィッティング

フィッティングする方法として，本研究では以下のような Jensen–Shannon (JS) 情報量を利用する．

$$D_{\mathrm{JS}}(p_{\mathrm{O}}, q) = \frac{1}{2}D_{\mathrm{KL}}(p_{\mathrm{O}}, m) + \frac{1}{2}D_{\mathrm{KL}}(m, p_{\mathrm{O}}).$$

ここで p_{O} は観測データを基にした頻度分布における確率密度値 $[0,1]$, q 提案モデル式から生成された頻度分布における確率密度値 $[0,1]$，D_{KL} は Kullback–Leibler (KL) 情報量，そして m は $\frac{1}{2}(p_{\mathrm{O}} + q)$ である．もし $p_{\mathrm{O}} \simeq q$ ならば，D_{JS} の値はゼロに近づく．パラメータ μ および σ の値は D_{JS} の最小値となるように決定される．

しかしながら，電子メールサイズ頻度の真の分布は不明である．本研究で行うことは観測データに基づく頻度分布およびモデルに基づく頻度分布とのフィッティングであることから，それらの確率密度を相対頻度に置き換えた上で JS 情報量の式を利用する．その際，モデル式 1 のパラメータ a の値はフィッティングする頻度分布に応じて調整される．

観測値は離散値であることから，KL 情報量は離散系である

$$D_{\mathrm{KL}}(p_{\mathrm{O}}, m) = \sum_s p_{\mathrm{O}} \ln \frac{p_{\mathrm{O}}}{m},$$
$$D_{\mathrm{KL}}(m, p_{\mathrm{O}}) = \sum_s m \ln \frac{m}{p_{\mathrm{O}}}.$$

を用いる．そして $D_{\mathrm{JS}}(p_{\mathrm{O}}, q)$ は観測値の範囲で用いられる．観測値の範囲の中にはいくつかのサイズの欠損があるが，それら欠損は観測値全体の中では軽微であると考えている．

モデル式 $p(s)$ のパラメータ μ および σ を決定する方法としては最尤法が存在し，μ および σ は $\mu = \frac{1}{N}\sum_i \ln \ln s_i$ および $\sigma = \frac{1}{N}\sum_i (\ln \ln s_i - \mu)^2$ と解析的に求まる．ここで s_i は各電子メールサイズ，N は観測データ数である．しかしながら，最尤法を用いたフィッティングでは良好な結果を示さなかったことから，我々は JS 情報量による方法を代用的に用いた．

表 2 図 4，図 5，図 6，そして 図 7 での各パラメータ値．
Table 2 Parameter values for each subdistribution in Figures 4, 5, 6, and 7.

	$p_{\mathrm{P}}(s)$	$p_{\mathrm{H}}(s)$	$p_{\mathrm{T}}(s)$	$p_{\mathrm{B}}(s)$
μ	1.0951	1.5371	85.0267	2.0443
σ	0.1969	0.1937	6.4375	0.2495

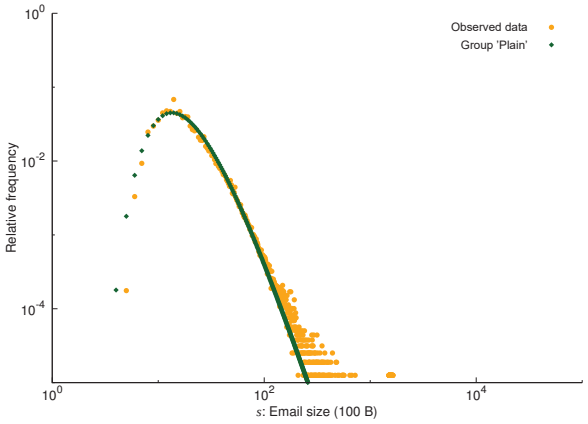

図 4 グループ ’Plain’ の頻度分布へのフィッティング結果．
Fig. 4 The fitted result of our proposed model for group “Plain.”

実際にフィッティングした結果を **図 4**，**図 5**，**図 6**，そして **図 7** にそれぞれ示す．フィッティングの際のパラメータは **表 2** に示す．各図は相対頻度の形で表していることに注意である．

図 4 で “Plain” は $s = 14$ 以上の領域でべき則性を有しているように読み取れる．このフィッティングに対する相関係数値を算出してみたところ，0.9861 になる．

図 5 で “HTML” は $s = 60$ 以上の領域でべき則性を有しているように読み取れる．相関係数値は 0.9507 になる．

図 6 で ‘Text attachment” は全体の形に不鮮明さが読み取れるが，本研究では今回得られたデータを基に分析する．この不鮮明さを解消するためには更なる送信記録収集が必要と考えられる．この図からは $s = 30$ 以上の領域でべき則性を有しているように考えられる．相関係数値は 0.7306 になる．

図 7 で “Binary attachment” は $s = 300$ 以上の領域でべき則性を有しているように考えられる．併せて，“Binary attachment” は $s = 100$ でピークを読み取れる．このピークの原因は業務用の特定のメールアドレス 1 つからの送信であった．相関係数値は 0.6351 になる．

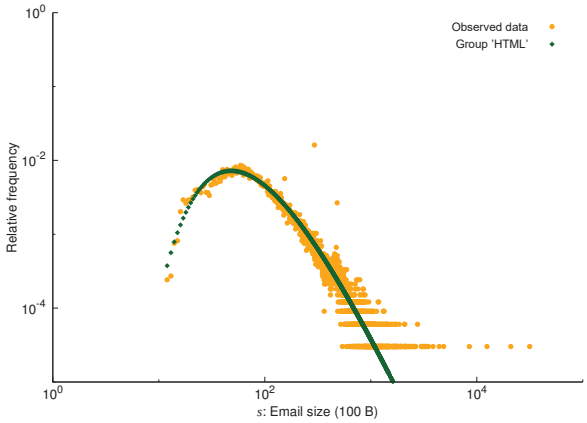

図 5　グループ 'HTML' の頻度分布へのフィッティング結果.

Fig. 5　The fitted result of our proposed model for group "HTML."

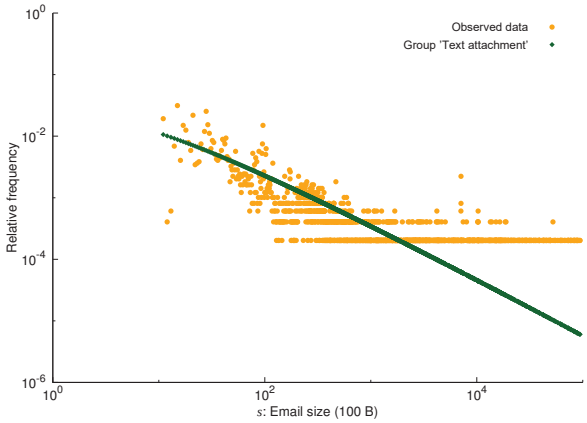

図 6　グループ 'Text attachment' の頻度分布へのフィッティング結果.

Fig. 6　The fitted result of our proposed model for group "Text attachment."

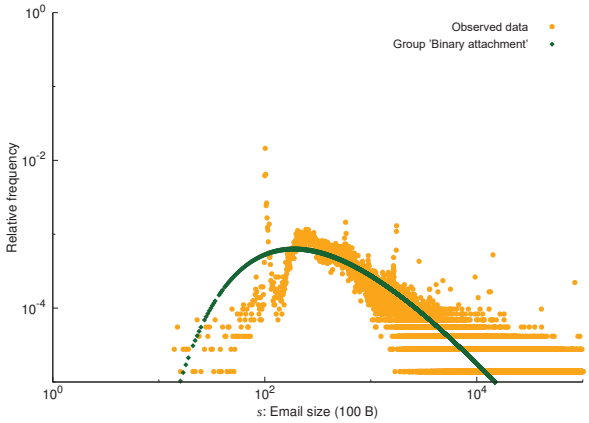

図 7　グループ 'Binary attachment' の頻度分布へのフィッティング結果.

Fig. 7　The fitted result of our proposed model for group "Binary attachment."

4.2　べき則性の評価

提案モデル (式 1) のべき則性を評価する方法として，我々は先行研究 [12] で用いられた方法を採用する．その方法は，電子メールのサイズの頻度分布を基に そのサイズの上限値を大きくしながら長期相関を有する疑似データ流量の時系列データを作成し，それら時系列データの長期相関を評価することによりサイズの頻度分布のべき則性を間接的に評価する．これはべき則性を直接評価する良好な方法を見い出せていないことから採用した方法である．

評価の手順は次の通りである．疑似サイズ 1 つは，電子メールの送信リクエスト 1 回毎に頻度分布から採択棄却法 (acceptance–rejection method) に基づきランダムに作成される．疑似データ流量の時系列データは，送信時刻毎にリクエスト回数分の疑似サイズの合計値を求め，それを送信期間分用意することにより作成される．この考えを**図 8** に示す．送信リクエスト回数の時系列データは長期相関を有するものが必要である．先行研究 [12] での実験の結果，もし電子メールサイズの頻度分布がべき則に従う場合，疑似データ流量の時系列データの長期相関は電子メールサイズの上限を大きくするに従いランダムに近づくことが見出されている．そのために長期相関を有する送信リクエスト回数の時系列データを事前に用意することになる．疑似データ流量の時系列データの長期相関は DFA [20,21] を用いることにより その程度が算出される．同一の頻度分布に対してサイズの上限をいくつか変更した疑似データ流量の時系列データを用意して，それらの長期相関の程度を DFA で算出する．もし電子メールサイズの頻度分布がべき則に従う場合，電子メールサイズのより大きな上限に対する疑似データ流量の時系列データの DFA 計算値はランダムを意味する 0.5 へ近づく．

今回用いる送信リクエスト回数の時系列データは，先行研究 [12] で収集された実際の送信リクエスト回数の時系列データである．その時系列データの期間は 2008 年 5 月 9 日から 2011 年 5 月 8 日であり，その期間での電子メールの送信リクエスト回数は合計 2,480,943，時間単位は 1 時間である．同じ時系列データを採用することにより先行研究 [12] での評価結果と比較しやすくしている．その送信リクエスト回数の時系列データは 1 週間周期を含んでいることから，疑似データ流量の時系列データにも 1 週間周期は含まれている．その周期性は先行研究と同様に除去され，長期相関の程度は その除去された疑似データ流量の時系列データから DFA により求められる．

実際に疑似データ流量の時系列を作成して DFA により長期相関の程度を計算した結果を**図 9** に示す．元の送信リクエスト回数の時系列データ対する DFA の計算値 α は 約 0.76 である．提案モデル式 1 で上限値を 10kB ($s = 100$) に設定した場合は $\alpha \approx 0.76$ となり，これは元の送信リクエスト回数の時系列データの場合と同じである．図 9 からは，電子メールサイズの上限値を大きくするにつれ α の値は 0.5 つまりランダムへ近づくことを読み取れる．このことから，提案モデル式はべき則性を有することが示される．

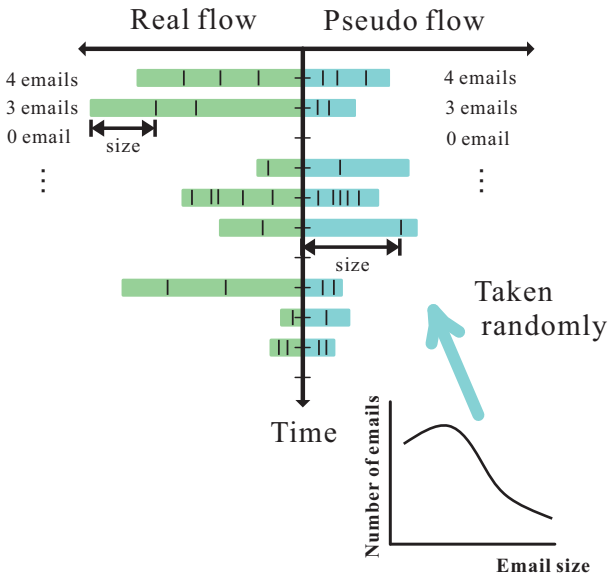

図 8　疑似データ流量を作成する仕組み.

Fig. 8　Schematic of the construction of an artificial flow.

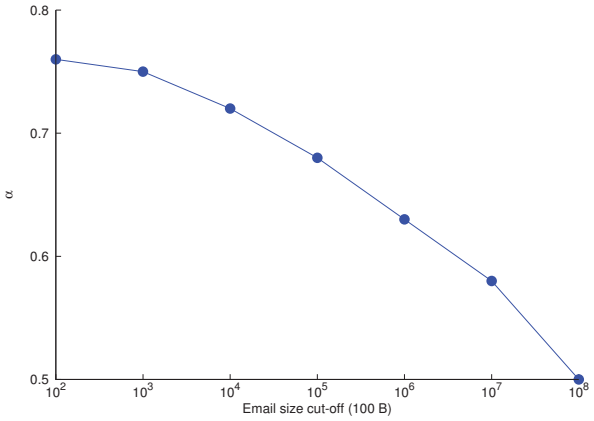

図 9　提案モデル式に基づく疑似データフローに対する DFA 分析結果.

Fig. 9　The DFA result of pseudo data flow used our proposed model.

5. まとめ

電子メール送信サーバで処理された電子メールのサイズの頻度について，それを説明するためのモデルを提案した．その頻度分布がべき則に従う場合，送信データ流量の時系列データの長期相関は送信リクエスト回数の時系列データの長期相関よりもランダムに近くなることが分かっている [12]．しかしながら，べき則を生じる仕組みは完全には解明されていない．

始めに，電子メールサイズの頻度分布形を詳細に分析するために，サイズの区間幅を 1kB から 100 バイトへ変更した．この変更により，1kB よりも小さいサイズの電子メールも考慮したより詳細な頻度分布形を得ることができた．

観測データから得られる電子メールのサイズ頻度分布の形からは，その形は異なる種類の電子メールのサイズ頻度分布の合成であるように考えられる．そこで，我々は電子メールの種類を表す電子メールヘッダの 1 つである “Content-Type” に注目し，“Content-Type” の内容を基に 4 種類の電子メール (平文，HTML，テキストファイル添付，バイナリファイル添付) を定義した．その定義に従い，我々はサイズ頻度分布を電子メールの種類毎に分解した．

構成要素に分解された各頻度分布を説明するためのモデルを提案した．その式は，対数正規分布 $f(x)$ および $x = \ln s$ の合成である．ここで，x はメール本文中の新規作成文の長さ，s は電子メールのサイズである．提案モデルでは，電子メールのヘッダ部，返信メールにおける引用文，HTML タグ，そして添付ファイルは x の余剰と扱われ，その関係を $x = \ln s$ で表す．そして，$f(x)$ 中の $\ln x$ は x の桁と見なされる．$f(x)$ は対数正規分布であることから，$\ln x$ は正規分布に従うことになる．このことから，電子メールの送信者は意識的もしくは無意識的に新規作成文の桁を正規分布に従い管理していることになる．

構成要素に分解された各頻度分布へ提案モデルのフィッティングを行い，概ね受け入れられる結果となった．

我々は，提案モデルがべき則を有することの評価を行った．その方法は，提案モデルを基に送信リクエスト毎に生成した電子メールサイズから疑似的な送信データ流量の時系列データを作成し，その長期相関の程度を計算するものである．先行研究 [12] では，電子メールサイズの上限値を大きくすることでその長期相関はランダムに近づくことが実験により示されている．我々は，その実験を提案モデルに適用する．評価の結果，提案モデルを基に疑似的に作成された送信データ流量の時系列データの長期相関は，その実験結果に従うものであった．これにより，提案モデルがべき則を有することが示された．

$f(x)$ は その式の構造上，両対数グラフ上ではべき則を意味する直線ではなく 2 次曲線として近似できる．従って，提案モデルのべき則性は $x = \ln s$ から生じることになる．$x = \ln s$ の関係を確認するためには，x つまり電子メール本文中の新規作成文の長さを分析することが要求される．もし x の値が分かれば，その頻度分布が $f(x)$ に従うか否かも判明する．しかしながら，今回データを収集した組織においては，電子メールに記述されている文章は個人情報保護の対象であることから，その分析は将来課題である．より良いモデル構築のために，今後も送信データの収集を続け，新たな視点が見つかれば そのためのデータ収集の許可を得られるように努めていく．

謝辞　データの収集をご快諾いただき，分析に際してご協力いただきました佐賀大学総合情報基盤センター各位に感謝します．

参考文献

[1] Faloutsos, M., Faloutsos, M. P. and Faloutsos, C.: On Power–Law Relationship of the Internet Topology, *Proceedings of the ACM SIGCOMM*, Vol. 29, pp. 251–262 (1999).

[2] Paxson, V. and Floyd, S.: Wide area traffic: the failure of Poisson modeling, *IEEE/ACM Trans. Networking*, Vol. 3, pp. 226–244 (1995).

[3] Csabai, I.: 1/f noise in computer network traffic, *Journal of Physics A: Mathematical and General*, Vol. 27, No. 12, p. L417 (online), available from ⟨http://stacks.iop.org/0305-4470/27/i=12/a=004⟩ (1994).

[4] Takayasu, M., Takayasu, H. and Sato, T.: Critical behaviors and $1/f$ noise in information traffic, *Physica A*, Vol. 233, pp. 824–834 (1996).

[5] Tadaki, S.: Power-Law Fluctuation in Internet Traffic, *Journal of the Physical Society of Japan*, Vol. 76, No. 3, pp. 044001–044001–5 (2007).

[6] Eckmann, J. P., Moses, E. and Sergi, D.: Entropy of dialogues creates coherent structures in e-mail traffic, *Proceedings of the National Academy of Sciences*, Vol. 101, No. 40, pp. 14333–14337 (online), available from ⟨http://www.pnas.org/content/101/40/14333⟩ (2004).

[7] Barabási, A. L.: The origin of bursts and heavy tails in human dynamics, *Nature*, Vol. 435, pp. 207–211 (2005).

[8] Goh, K. I. and Barabási, A. L.: Burstiness and memory in complex systems, *EPL (Europhysics Letters)*, Vol. 81, No. 4, p. 48002 (online), available from ⟨http://stacks.iop.org/0295-5075/81/i=4/a=48002⟩ (2008).

[9] Malmgren, R. D., Stouffera, D. B., Motter, A. E. and Amaral, L. A. N.: A Poissonian explanation for heavy tails in e-mail communication, *Proceedings of the National Academy of Sciences*, Vol. 105, No. 47, pp. 18153–18158 (online), available from ⟨http://www.pnas.org/content/105/47/18153⟩ (2008).

[10] Anteneodo, C., Malmgren, R. D. and Chialvo, D. R.: Poissonian bursts in e-mail correspondence, *The European Physical Journal B*, Vol. 75, pp. 389–394 (online), available from ⟨http://www.springerlink.com/content/t1321475062jm273/⟩ (2010).

[11] Karsai, M., Kaski, K., Barabási, A. L. and Kertész, J.: Universal features of correlated bursty behaviour, *Scientific Reports*, Vol. 2 (online), available from ⟨http://dx.doi.org/10.1038/srep00397⟩ (2012).

[12] Matsubara, Y., Hieida, Y. and Tadaki, S.: Fluctuation in e-mail sizes weakens power-law correlations in e-mail flow, *The European Physical Journal B*, Vol. 86 (online), available from ⟨http://dx.doi.org/10.1140/epjb/e2013-40209-x⟩ (2013).

[13] 松原義継，武藏泰雄：送信された電子メールサイズの頻度に現れるべき則の分析，インターネットと運用技術シンポジウム 2014 論文集，Vol. 2014, pp. 71–77 (2014).

[14] Freed, N. and Borenstein, N.: Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies, RFC 2045 (Draft Standard) (1996). Updated by RFCs 2184, 2231, 5335, 6532.

[15] Freed, N. and Borenstein, N.: Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types, RFC 2046 (Draft Standard) (1996). Updated by RFCs 2646, 3798, 5147, 6657.

[16] Moore, K.: MIME (Multipurpose Internet Mail Extensions) Part Three: Message Header Extensions for Non-ASCII Text, RFC 2047 (Draft Standard) (1996). Updated by RFCs 2184, 2231.

[17] Freed, N. and Borenstein, N.: Multipurpose Internet Mail Extensions (MIME) Part Five: Conformance Criteria and Examples, RFC 2049 (Draft Standard) (1996).

[18] Freed, N. and Klensin, J.: Multipurpose Internet Mail Extensions (MIME) Part Four: Registration Procedures, RFC 4289 (Best Current Practice) (2005).

[19] Freed, N., Klensin, J. and Hansen, T.: Media Type Specifications and Registration Procedures, RFC 6838 (Best Current Practice) (2013).

[20] Peng, C. K., Buldyrev, S. V., Havlin, S., Simons, M., E, S. H. and Goldberger, A. L.: Mosaic organization of DNA nucleotides, *Phys Rev E*, Vol. 49, No. 2, pp. 1685–1689 (1994).

[21] Peng, C. K., Havlin, S., Stanley, H. E. and Goldberger, A. L.: Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series, *Chaos*, Vol. 5, pp. 82–87 (1995).