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

松原 義継[1,a)]　武藏 泰雄[2,b)]

概要：とある大学で利用者の異なる 2 種類の電子メール送信サーバで処理された電子メールサイズの頻度分布を年度単位毎に分析した．分析に用いた各電子メール送信サーバでの送信数は，それぞれ 450 万 4760 通 (5 年間分) および 87 万 7271 通 (6 年間分) である．頻度分布の大きな形からは，両方共にべき則を読み取れる．頻度分布形にべき則が現れることについては，電子メール送信サービスの実運用において送信可能な電子メール 1 通のサイズに上限値を設けることに意味があることが示されていると考えられる．さらに，この頻度分布形を説明するためのモデル式の構築を試み，良好なシミュレーション結果を得られた．

キーワード：電子メールサイズ，べき則分布，ガンマ分布

Analysis of power-law distribution of e-mail sizes processed in e-mail sending servers

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

Abstract: Previously, we reported the power-law distribution in the frequency of e-mail sizes in a university campus network [1]. We analyzed the several distributions obtained from two e-mail servers in the university campus network. The numbers of send-requests are 4,504,760 e-mails (April 1st, 2009 - March 31st, 2014) and 877,281 e-mails (April 1st, 2008 - March 31st, 2014). The obtained distribution forms are a power-law distribution. We consider that these power-law distributions denote significance to set an upper limit for sending e-mails in actual e-mail services. We propose a model that explains the distribution, especially focused on the power-law distribution and it can explain the principal fluctuations.

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

1. はじめに

インターネットは，現代社会において重要な通信媒体の１つである．近年，その構造や流量に関する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が見出されている [2]．その他には，パケット流量の時系列上に見出されるべき則に従う相関，電子メールの送信間隔に見出されるべき則に従う相関等がある [3–12]．

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を分析したものがある [1]．その頻度分布形からはべき則を読み取ることができる．これは，電子メールのサイズおよびそのサイズを有する電子メールの頻度との間にべき乗の関係があることを示している．さらに その分布形に従う場合，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．

本論では，文献 [1] での分析結果を基に電子メールサイズの頻度分布に対する更なる分析を行う．電子メールのサイズの分布に関する先行研究は，我々の知りうる限り見当たらない．

分析に用いる電子メールサーバは，文献 [1] で用いられ

[1] 佐賀大学
Saga University, 1 Honjo-machi, Saga-shi, Saga, 840-8502, Japan
[2] 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

た電子メール送信サーバの他に利用者の異なる電子メール送信サーバを加えた 2 種類である．このように用意することにより，利用者が異なる場合での頻度分布形に共通性が現れるか否かを分析できる．分析に用いる電子メール送信数は，前者は 450 万 4760 通 (5 年間分)，後者は 87 万 7271 通 (6 年間分) である．

さらに，本論では頻度分布形の時間変化も分析する．文献 [1] で示された電子メールサイズの頻度分布は，2008 年 5 月から 3 年間分のデータを基に作成されているが，その頻度分布形の時間変化については分析されていない．本論では，頻度分布形の時間変化を分析することにより，べき則を読み取れるその頻度分布形が一過性のものであるか持続性のあるものであるかを検証する．

これら分析結果を踏まえ，本論ではその頻度分布形を説明するためのモデル式の考察およびそのモデル式のシミュレーションを行う．

利用者が電子メールを送信する理由は，各利用者の時々の事情により様々である．そ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. 送信データの取得

本論で扱う電子メールは，佐賀大学の学内ネットワーク (ネットワークアドレス 133.49.0.0/16) 内端末群から同大学総合情報基盤センターにてサービス中である電子メール送信サーバに送信され処理されたものである．送信可能な電子メール 1 通のサイズの上限値である 10 メガバイト [MB] を超える電子メールは集計に入らない．電子メール送信サーバは，教職員用および主に学生が利用する教育用の 2 種類がある．本論では これら 2 種類の送信サーバのログファイルを それぞれ分析する．ログファイルの期間は，教職員用は 2009 年 4 月 1 日から 2013 年 3 月 31 日までの 5 年間分，教育用は 2008 年 4 月 1 日から 2013 年 3 月 31 日までの 6 年間分である．

教職員用の送信数の時系列データを図 1 に示す．縦軸は電子メール数，横軸は年月，集計単位は 1 ヶ月間である．図 1 からは，2009 年度付近のデータが他の年度よりも高めに読み取れる．詳細を調べたところ，これは大学の教務システムから学生に対する連絡メールの影響であることが分かった．2010 年度中からは教務システム側の仕様変更に伴い電子メールの送信方法が変更されているため，2010 年度中以降は教務システムの影響が読み取れなくなっている．本論では，教務システムからの送信記録は分析対象から除外する．その分を除外した時系列データを図 2 に示す．本論では，教職員用の分析として図 2 での各電子メールを用いる．

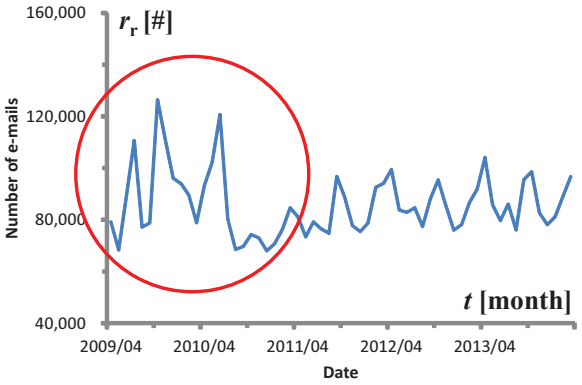

図 1 教職員用電子メール送信サーバで処理された電子メール数．
Fig. 1 Time series of e-mails in the e-mail server for staffs.

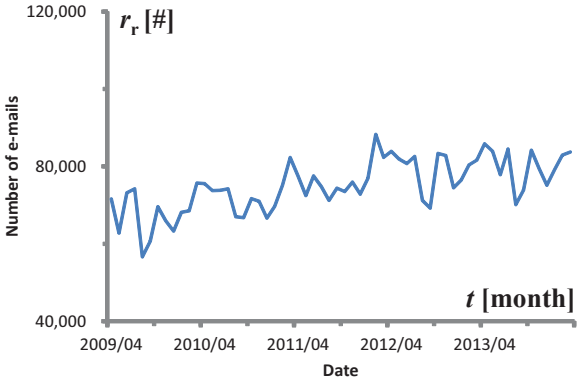

図 2 分析に用いる教職員用電子メール送信サーバで処理された電子メール数．
Fig. 2 The time series which is analyzed in this paper.

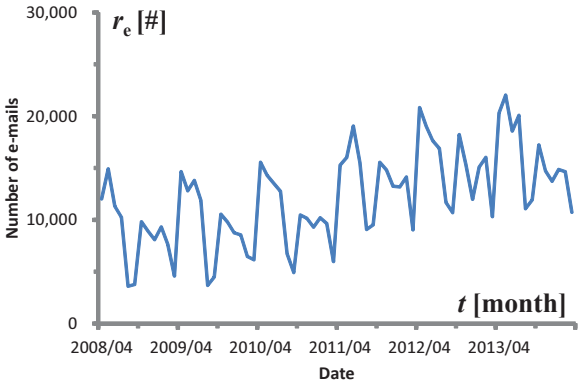

図 3 教育用電子メール送信サーバで処理された電子メール数．
Fig. 3 Time series of e-mails in the e-mail server for students.

教育用の送信数の時系列データを図 3 に示す．縦軸は電子メール数，横軸は年月，集計単位は 1 ヶ月間である．全体として送信回数は増加傾向であり，図 3 からは 1 年間周期を読み取れる．

3. 分析

前節で得られた教職員用の図 2 および教育用の図 3 を形成している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，学生は新入生が入学する．これは電子メールの利用者が毎年 4 月に大きく変更されることを意味し，この変更に伴い電子メールの利用方法に影響が現れると考えられる．この事を鑑みて，本論では年度単位で集計する．本論では，サイズの単位はキロバイト [KB] である．

作成した頻度分布を教職員用および教育用でそれぞれ図 4 および図 5 に示す．両図内での表記 ‘AY’ は 学年度 (academic year) の略記である．各図の縦軸は電子メールサイズの頻度を該当年度の電子メール総数で割った値，横軸は電子メールサイズ (キロバイト単位) である．両軸とも対数化されていることに注意である．図中の各青点は実際の値を表す．赤線は図全体の傾きを表している．傾きは青点全体を最小二乗法でフィッティングすることで得られている．全体の傾きが分かるように赤線の引かれている位置は全体的に上にずらしている．

3.1 頻度分布の形

図 4 および図 5 に描かれている各年度の分布は，一部に形の歪みはあるが，その大きな形として直線的であるように読み取れる．図 4 および図 5 中の各頻度分布は両対数グラフであり，両対数グラフ上で直線的であることは，その頻度分布の形はべき則に従っていることを意味している．

3.2 頻度分布形の実運用上の意味

電子メール送信サービスを実運用する観点から，電子メールサイズの頻度分布形がべき則として読み取れることからは 1 つの実用的な知見を得ることが可能である．それは，送信可能な電子メール 1 通のサイズに上限値を設けることに意味があることである．電子メール送信サービスの実運用上，電子メール送信サービスを不安定化させる程に大きなサイズの電子メールを処理させることは望ましくない．その対策としては，送信可能な電子メール 1 通のサイズに上限値を電子メール送信サーバに設定する方法がある．

もし，実際の電子メールサイズの頻度分布形が正規分布のような指数関数系の分布である場合，電子メール送信サービスを不安定化させる程に大きなサイズの電子メール送信は確率的に滅多に起きない可能性がある．そのような場合，サイズの上限値を設けなくとも実運用上の支障は少ない可能性が出てくる．

本論での分析からは，その頻度分布形の大きな形はべき則として読み取れることから，指数関数系の分布では滅多に生じない大きなサイズの電子メールの送信は比較的に高確率で発生することを意味する．このような場合，上限値を設けることはサービスの安定運用の観点から望ましい．

3.3 頻度分布形の詳細

図 4 における 2009 年度および 2010 年度の頻度分布からは，電子メールサイズ 約 15 KB から 40 KB 付近で形の歪みを読み取れる．図 5 からは全ての年度で約 15 KB から 40 KB 付近で形の歪みを読み取れる．このような形を生じる原因として，頻度分布形は電子メールの主たる中身が文章の場合および文章以外のものが添付されている場合で異なっており，全体としての形は図 4 および図 5 に示されている形であると考えられる．

実際に電子メールを作成する場合，文章のみで数十 KB を超え数 MB に至るまでの電子メールを作成することは困難である．本論では，頻度分布は電子メールの中身が文章主体で構成されるサイズ領域および文章以外のものを主体として構成されるサイズ領域という 2 つのサイズ領域で分割可能と考える．そこで，図 4 および図 5 中の各頻度分布を以下のようなサイズ領域で分割する．

(1) 電子メールサイズが 15 KB 未満領域.
電子メールの中身の主たる構成要素は文章であり，そのサイズは文章のサイズに影響される．電子メールの形式として HTML 形式メールがある．HTML 形式メールの場合は，HTML タグの量にも影響される．

(2) 電子メールサイズが 15 KB 以上 10 MB 以下の領域.
この領域では，表計算ソフトやグラフ作成ソフト等により作成されるファイルを電子メールに添付することによるサイズへの影響が大きい．例えば，デジタルカメラで撮影した画像ファイル 1 つを電子メールに添付することにより その電子メールのサイズは数 MB になることも起こりえる．

領域を 15 KB を境に 2 分割した理由は，次の手順に基づく実験結果に基づいている．

(1) 空文の表計算ファイル，空文の PDF ファイル，空文の文章ファイル，空文のプレゼンテーションファイルを それぞれ用意する．

(2) これら各ファイルを添付した空文の電子メールをぞれぞれ作成して送信する．作成の際，電子メールの件名も空文にする．

(3) 受信した各電子メールのサイズを見る．

実験に用いたメールソフトは Windows 版 Mozilla Thunderbird 31.2.0 である．表計算ソフト，文章作成ソフト，プレゼンテーションソフトは，マイクロソフト社 Office 2010 での各ソフトである．各電子メールのサイズは，それぞれ 14 KB，20 KB，20 KB，そして 40 KB であった．これら

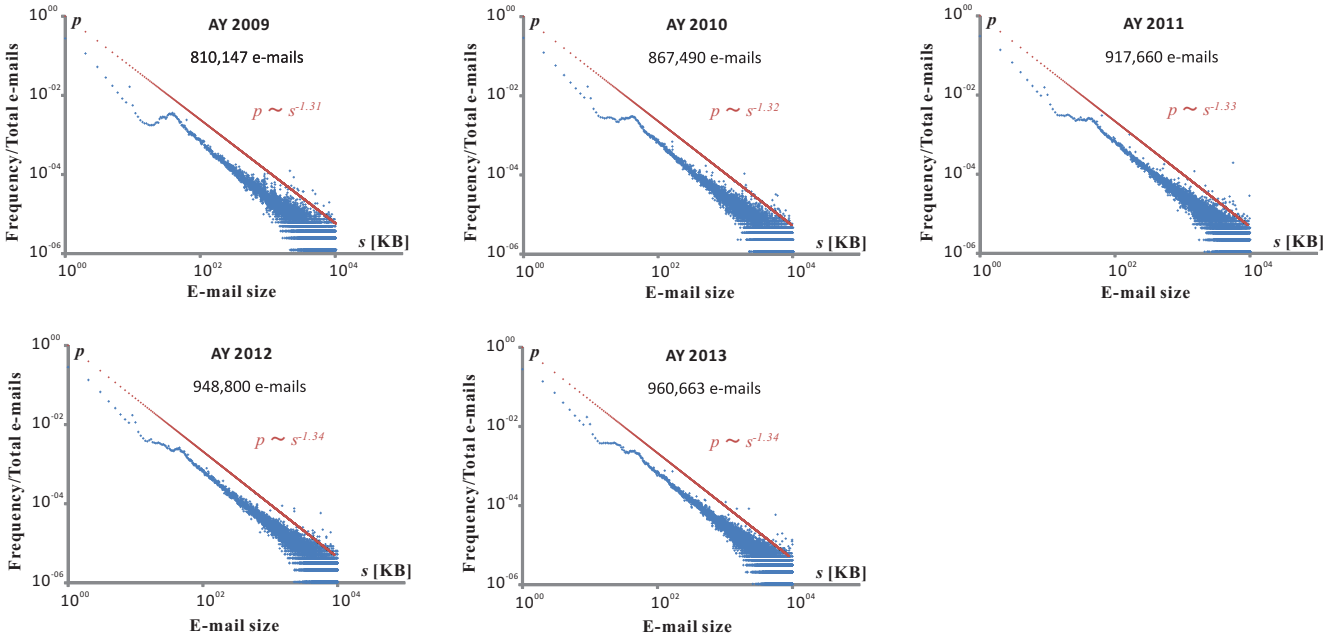

図 4　教職員用電子メール送信サーバにおける電子メールサイズ頻度の年度別変化.

Fig. 4　Frequency in e-mail sizes for staffs per academic year.

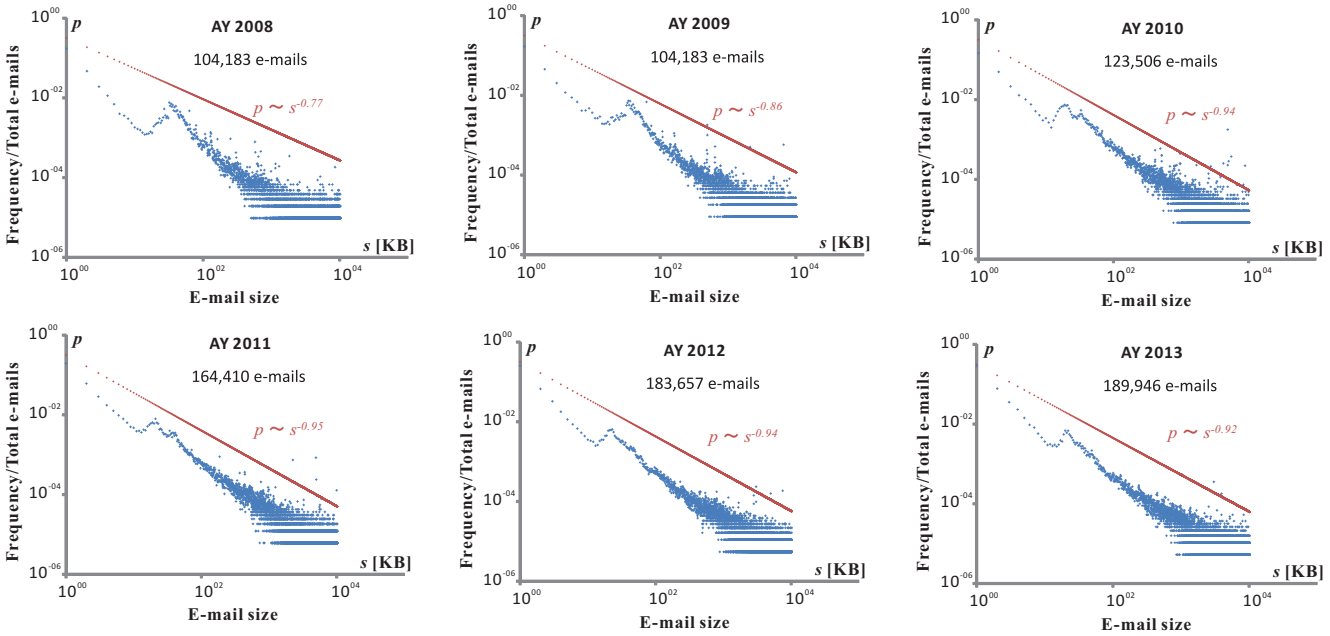

図 5　教育用電子メール送信サーバにおける電子メールサイズ頻度の年度別変化.

Fig. 5　Frequency in e-mail sizes for students per academic year.

の結果から，領域を分ける目安として 本論では 15 KB を設定した.

3.4　頻度分布形の時間変化に対する分析

図 4 および図 5 からは，年度の経過に伴い全体の形はより直線的なものに変化しているように読み取れる．考えられることとしては，文章主体で構成される電子メール送信が増加していることである．例えば，近年普及してきている グループウェア，ネットワークストレージのような媒体でファイルを送信していることが考えられる．この場合の電子メールの役割はファイルの置き場所を相手に伝えることと考えられるため，電子メールサイズは比較的小さくなっている可能性がある．さらに，HTML 形式の電子メールの普及に伴い，電子メールの平均的なサイズが変化している可能性もある.

4.　モデル式構築のための考察

図 4 および図 5 に示されている頻度分布形を表すことの可能なモデル式について考察する.

著者らが注目したのは，自らの経験則である，論文もし

くはレポートの文章が長文になる程，図もしくは写真のような文章ではないものが含まれ易いことである．さらに，文章ではないものが含まれている電子メールサイズは，文章のみの電子メールサイズよりも明らかに大きい場合が多い．これらの経験則を基に図 4 および図 5 中の頻度分布形を説明可能なモデル式を考察した．

4.1 モデル式の定義

考察の結果，文章のみで構成される電子メールサイズの分布はガンマ分布であるとみなし，その分布式に経験則を式化したものを組み込むことにより説明できる可能性があることが分かった．電子メールではない場合，言語学の研究成果の中に 1 文の長さの頻度分布を分析した先行研究が存在し，それら先行研究からは数種類の頻度分布が報告されている [13–17]．しかしながら，電子メールについては，我々の知りうる限り先行研究は見当たらない．言語学での先行研究で報告されている頻度分布の中にガンマ分布 [14] およびガンマ分布ライクな分布 [16] が存在する．ガンマ分布を基にモデル式を考察したところ，(1)1 文から複文への拡張が理論的に可能，(2) ガンマ分布の式中にべき則が含まれていることからべき則の再現が容易，なことから今回のモデル式の構築に至った．このことを整理したものを定義として以下に示す．

(1) 文章の 1 文の長さの頻度分布はガンマ分布に従う．ここで，1 文の長さは文字数である．もし添付ファイル中に文章が含まれている場合は，その 1 文の長さの頻度分布形もガンマ分布に従う．

(2) 複数の文からなる複文の長さの頻度分布は，1 文の長さの頻度分布の和合成形であるガンマ分布とする．

(3) 先に述べた経験則の数式化として，電子メールサイズ s と複文の長さ x との間の関係を以下のように定義する．

$$s = x + a^x - 1 \quad (s \geq 0, x \geq 0, a > 1), \tag{1}$$

ここで，a^x は文章でないものの長さである．例えば，図，写真，音声等である．文章の長さである x と文章でないものの長さとの関係は指数関数的であると定義し，定数 a を導入することにより そのことを簡潔に表している．右辺第 3 項目の -1 は，$x = 0$ の際に $s = 0$ となるための補正項である．

(4) モデル式は，式 1 を x に関する頻度分布式であるガンマ分布 $g(x) = (1/C)\lambda^\alpha x^{\alpha-1}\exp(-\lambda x)$ に組み込むことにより s に関する頻度分布式として得られる．
この際，式 1 はランベルトの W 関数を用いることにより，以下のように変形可能である．

$$x = \frac{1}{\ln a}\ln\frac{W(a^{s+1}\ln a)}{\ln a}. \tag{2}$$

ここで，$W()$ はランベルトの W 関数である．

式 2 を $g(x)$ に組み込むことで得られるモデル式 $p(s)$ は以下のようになる ($s > 0$ KB).

$$p(s) = \frac{1}{C}\beta^\alpha\{\ln\frac{W(a^{s+1}\ln a)}{\ln a}\}^{\alpha-1} \frac{\ln a}{1+W(a^{s+1}\ln a)}\{\frac{W(a^{s+1}\ln a)}{\ln a}\}^{-\beta}, \tag{3}$$

$$\beta = \frac{\lambda}{\ln a}, \tag{4}$$

ここで C は正規化定数，$\beta > 0, \alpha > 1, \gamma > 0, a > 1$, そして $W()$ はランベルトの W 関数である．

4.2 モデル式の性質

$x \gg 1$ の場合，式 1 は，$s \sim a^x$ と近似される．この場合の $p(s)$ は

$$\begin{aligned} p(s) &\sim \lambda^\alpha(\frac{\ln s}{\ln a})^{\alpha-1}(s\ln a)^{-1}\exp(-\lambda\frac{\ln s}{\ln a}) \\ &\quad \text{(Using } \beta \text{ as defined in equation 4.)} \\ &= \beta^\alpha(\ln s)^{\alpha-1}s^{-1}\exp(-\beta\ln s). \end{aligned} \tag{5}$$

両辺を対数化すると

$$\begin{aligned} \ln p(s) &\sim (\alpha-1)\ln(\ln s) - \ln s - \beta\ln s \\ &= (\alpha-1)\ln(\ln s) - (\beta+1)\ln s \\ &\quad (\text{where } \ln(\ln s) < \ln s) \\ &\sim -(\beta+1)\ln s. \end{aligned} \tag{6}$$

従って

$$p(s) \sim s^{-(\beta+1)}. \tag{7}$$

このように，モデル式は $x \gg 1$ の領域ではべき則として近似可能であることが分かる．

もし x の値が 0 KB に近い場合，式 1 は $s \sim x$ と近似可能である．この場合，$p(s)$ の形は ガンマ分布 $g(x)$ の形に近づく．

5. モデル式のシミュレーション

モデル式 $p(s)$ により生成される電子メールサイズの頻度分布形が実際の頻度分布形を近似可能であるか否かをシミュレーションで検証した．

本論では 1 例として教職員用である図 4 中の 2009 年度の頻度分布に対するシミュレーション結果を示す．今回のシミュレーションで用いるモデル式の各パラメータ値は以下の通りである．

(1) $s < 15$ KB: $a = 1.01, \alpha = 1.01, \beta = 24.$

(2) 15 KB $\leq s \leq$ 10 MB: $a = 1.1, \alpha = 3, \beta = 0.5.$

規格化定数 C は以下の式により求めた．$C = \sum_{s=1}^{10\ \mathrm{MB}} p(s)$.

シミュレーション結果を図 6 に示す．図 6 中の各軸および青い点は図 4 の場合と同じである．図 6 中の赤線はモデ

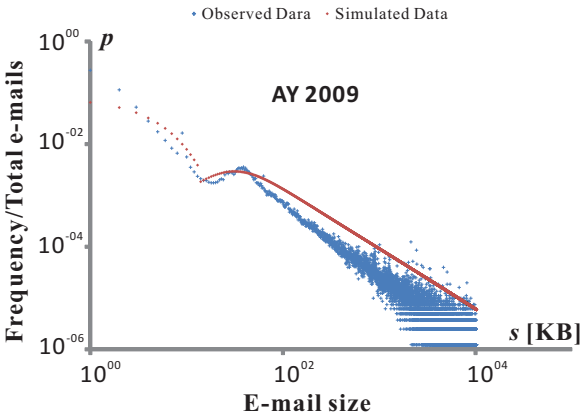

図 6　モデル式のシミュレーション結果．比較対象の頻度分布は図 4 の 2009 年度の頻度分布．

Fig. 6　The simulated result of our proposed model. Compared distribution is academic year 2009 for staffs.

ル式による計算結果である．シミュレーション結果は，実際の頻度分布形に近い形，特に大きなサイズ領域でのべき則の形を表せているように読み取れる．電子メールサイズが 10 KB より小さくなるとシミュレーション値は実際よりも小さいように見えるが，シミュレーションにより表せた形は実際の頻度分布の大きな形を概ね説明可能と考えられる．

6. 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分析を行った．分析に用いたデータは，とある大学で運用されている利用者の異なる 2 種類の電子メール送信サーバから得られた．それぞれの送信数は，450 万 4760 通 (5 年間分) および 87 万 7271 通 (6 年間分) である．頻度分布は，サイズ的にはキロバイト単位で時間的には年度単位で作成された．

得られた各頻度分布形は，その大きな形としてべき則を読み取れる．頻度分布形にべき則を読み取れることからは，電子メール送信サーバ側に処理可能な 1 通のサイズに上限値を設けることに意味があることが示されていると考えられる．サイズの頻度分布形がべき則の下で送信可能な電子メール 1 通のサイズに上限値が設定されていない場合，サービスを不安定化させる程に大きなサイズの電子メールの送信処理は 正規分布のような指数関数系の分布よりも比較的高い確率で発生する可能性がある．これは，そのサービスを提供する側にとってはサービスの質の低下を招く可能性がある．本論からは，送信可能な電子メール 1 通のサイズに上限値を設定することについて，実際のデータを基に統計的観点から知見を得ることができた．

頻度分布形の詳細には，形の一部分に歪みを読み取れる．この理由として，文章以外のものが添付されている場合の頻度分布形とそうでない場合の頻度分布形に違いがあるためと考えられる．

同時に その形の歪みは，年度の進行に伴い読み取りにくくなっている．これは，送信者側の需要の変化に伴い，電子メールの中身の構成に変化が生じていることが考えられる．例えば，ファイルの送信手段としてグループウェアもしくはネットワークストレージが普及していることに伴う電子メール経由でのファイル送信の頻度低下，HTML 形式電子メールの普及に伴う平均的なサイズの変化である．

このような頻度分布形を説明するためのモデル式の構築を試みた．モデル式は，ガンマ分布をベースに，グラフや写真のような文章ではないものの電子メールサイズへの影響を経験則として組み込んだ．シミュレーション結果を見る限り，このモデル式は実際の頻度分布の大きな形を概ね説明可能と思われる．もし，電子メールの本文の分析が可能ならば，詳細な分析および より正確なモデル式の考察が可能であるが，これらは今後の課題である．

参考文献

[1] 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).

[2] 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).

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

[4] 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).

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

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

[7] 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).

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

[9] 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).

[10] 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).
[11] 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).
[12] 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).
[13] 安本美典：文の長さの分布型について，計量国語学，pp. 20–24 (1958).
[14] 佐々木和枝：文の長さの分布型，計量国語学， Vol. 78, pp. 13–22 (1976).
[15] 新井皓士：文長分布の対数正規分布性に関する一考察：芥川と太宰を事例として，一橋論叢， Vol. 125, No. 3, pp. 205–223（オンライン），入手先 ⟨http://hermes-ir.lib.hit-u.ac.jp/rs/handle/10086/10418⟩ (2001).
[16] Sigurd, B., Eeg-Olofsson, M. and Van Weijer, J.: Word length, sentence length and frequency ? Zipf revisited, *Studia Linguistica*, Vol. 58, No. 1, pp. 37–52 (online), DOI: 10.1111/j.0039-3193.2004.00109.x (2004).
[17] 古橋　翔：文の長さ分布に見られる対数正規性，技術報告，国立国語研究所 第 1 回コーパス日本語学ワークショップ (2012).