Vol.2016-IOT-32 No.34
2016/3/4

DNSアクセスにおけるuniq率のエントロピーとの比較検討

松原 義継[1,2,a)]　武藏 泰雄[3,b)]

概要：DNS アクセスにおける外れ値検出方法として考案された uniq 率をエントロピーと比較検討した．本論文における uniq 率とは，単位時間当りの DNS サーバへのアクセス数に対するクライアント数もしくはクエリの種類数の割合である．uniq 率の定義上，与えられた uniq 率に対応するエントロピーには幅があることから，その程度について理論的に考察した．その結果，uniq 率の取り得る値の両端では uniq 率とエントロピーとの間にある値の幅は減少することが分かった．さらに，公開されているデータセットに基づき uniq 率およびエントロピーを算出し比較したところ，uniq 率の変動とエントロピーの変動との間には一定以上の相関が見られた．

キーワード：DNS ログデータ，uniq 率，エントロピー

A Comparative Study Of Uniq Rate and Entropy in DNS Access

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

Abstract: We compared the property of uniq rate with that of entropy in DNS access. The uniq rate is an outlier detection method which is a ratio of the number of types of clients or queries in the number of DNS access per unit time. Entropy has a range of the values corresponding to a value of uniq rate. Therefore, we considered the range theoretically. As a result, we found that the values of uniq rate close to that of entropy in the both ends of the possible values of uniq rate. Furthermore we compared timeseries of uniq rate and entropy in public datasets, we found a certain correlation between them.

Keywords: DNS log data, uniq rate, entropy.

1. はじめに

インターネットに代表されるコンピュータネットワーク上に存在する DoS (Deny of Services) のようなサービス不能攻撃やシステム上の脆弱性を悪用した攻撃は，コンピュータネットワークの社会的重要性の増加する状況において憂慮すべき事態である．これらの攻撃によるサービスへの悪影響を最小限に食い止めるための方法の 1 つとして外れ値検知があり，様々な方法が報告されている [1–7].

本論文では DNS アクセスに基づく外れ値検知方法として考案された uniq 率を同じく DNS アクセスに基づく外れ値検知方法であるエントロピーと比較検討する．uniq 率は，単位時間当りの DNS アクセス回数におけるアクセス元クライアント数もしくはクエリの種類数との割合である [8]．似たような方法としては，種類数のみで分析する‘異なり数’がある [9–12]．エントロピーは，単位時間当りにおけるアクセス元クライアント単位もしくはクエリ単位での DNS アクセスの偏りの程度を情報量の期待値を用いて数値化したものである [13–16].

DNS は，電子メールやウェブページのようなネットワークサービスにおける基本的なサービスであり，多くのネッ

[1] 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
[2] 佐賀大学
Saga University, 1 Honjo-machi, Saga-shi, Saga, 840-8502, Japan
[3] 熊本大学
Kumamoto University, 2-40-1 Kurokami Chuo-ku, Kumamoto-shi, Kumamoto, 860-8555 Japan
[a)] 146d9301@st.kumamoto-u.ac.jp
[b)] musashi@cc.kumamoto-u.ac.jp

トワークサービスから利用されている．そのため，DNS アクセスの分析からは，複数のネットワークサービスのアクセス動向を効率良く得ることが期待できる．

エントロピーを用いる方法は，DoS 攻撃や DDoS 攻撃 (Distributed Denial of Service attack) のような通常とは異なる偏りのあるアクセス動向を検知できる．uniq 率は，エントロピーよりも簡易な方法として考案された．もし uniq 率がエントロピーの簡易版としての実用性を認められるのであれば，uniq 率はエントロピーに基づく外れ値検知を行うか否かを判断する前処理としての利用が期待される．

本論文では，次の 2 点について uniq 率のエントロピーとの比較検討について述べる．

(1) uniq 率の定義上，uniq 率の値に対応するエントロピーの値には幅があることから，その幅の程度の理論的考察．
(2) インターネット上に公開されている 2 つのデータセットである ‘DARPA 2000’ および ‘CDX 2009’ に基づき，uniq 率の値の変動およびエントロピーの値の変動との相関の分析 [17,18].

2. uniq 率およびエントロピーの定義

本論文で用いる uniq 率およびエントロピーの定義を簡単に示す．

ある観測時間の区間 $[t, t+\Delta t)$ での確率変数を X_t，DNS アクセス回数を n_t とする．これ以後，観測時刻 t とは $[t, t+\Delta t)$ を意味する．

2.1 uniq 率

ある時刻 t におけるクライアント数もしくはクエリの種類数を m_t とする時，uniq 率 U_t を以下のように定義する．

$$U_t = \frac{m_t}{n_t}. \tag{1}$$

U_t の値の取り得る範囲は，その定義より

$$\frac{1}{n_t} \le U_t \le 1 \ (= \frac{n_t}{n_t}) \tag{2}$$

である．最小値である $1/n_t$ は，n_t 回全てのアクセス内容が同一 (1 種類) を意味する．一方，最大値である 1 は，n_t 回の各アクセス内容に重複のないこと (n_t 種類) を意味する．

本論文では，uniq 率 U_t の値の取り得る範囲を $[0,1]$ に正規化する．その理由は，U_t の取り得る値の下限は，観測時刻で異なる値を取り得る n_t に依存するためである．異なる観測時刻で U_t の値の取り得る範囲が異なると，異なる時刻での uniq 率の値同士を比較することに支障をきたす懸念がある．

そこで本論文で用いる uniq 率は，式 1 を基に

$$\begin{aligned} U'_t &= 1 + \log_{n_t} U_t \\ &= \log_{n_t} m_t \end{aligned} \tag{3}$$

とする．

本論文では，2 種類の uniq 率を用いる．それぞれを以下のように表す．

- $U'_{t,\mathrm{c}}$: クライアント数の内訳に基づく正規化された uniq 率
- $U'_{t,\mathrm{q}}$: クエリの種類数に基づく正規化された uniq 率

2.2 エントロピー

確率変数 X_t の値が i となる確率密度を p_i で表す時，X_t のエントロピー $H(X_t)$ は

$$H(X_t) = -\sum_{i \in X_t}^{m_t} p_i \log_2 p_i \tag{4}$$

である．

$H(X_t)$ の値の取り得る範囲は，

$$0 \le H(X_t) \le \log_2 n_t \tag{5}$$

である．最小値である 0 は，n_t 回全てのアクセス内容が同一 ($p_i = 1\ (= n_t/n_t)$) を意味し，$m_t = 1$ をも意味する．最大値である $\log_2 n_t$ は，各アクセス内容に重複のないこと ($p_i = 1/n_t$) を意味し，$m_t = n_t$ をも意味する．

本論文では，エントロピー $H(X_t)$ の値の取り得る範囲を $[0,1]$ に正規化する．その理由は，$H(X_t)$ の取り得る値の上限は，観測時刻で異なる値を取り得る n_t に依存するためである．異なる観測時刻で $H(X_t)$ の値の取り得る範囲が異なると，異なる時刻でのエントロピー値同士を比較することに支障をきたす懸念がある．そこで本論文では，取り得る範囲を正規化することにより，エントロピー値の取り得る範囲に対する観測時刻依存性をなくす．

本論文で用いるエントロピーは

$$H'_t = \frac{H(X_t)}{\log_2 n_t} \tag{6}$$

で表す ($0 \le H'_t \le 1$).

本論文では，正規化された 2 種類のエントロピーを用いる．それぞれを以下のように表す．

- $H'_{t,\mathrm{c}}$: クライアント数の内訳に基づく正規化されたエントロピー
- $H'_{t,\mathrm{q}}$: クエリの内訳に基づく正規化されたエントロピー

3. uniq 率の値に対応するエントロピー値の幅

ある時刻 t におけるクライアント数もしくはクエリの種類数 m_t の値が与えられる時，uniq 率の値はその定義により一意に求まる．一方，エントロピーの値は m_t の各要素 (各クライアントもしくは各クエリ) の内訳に複数の可能性がある．そのため，uniq 率の値に対応するエントロピー値には幅が存在し得ることになる．この幅を考察することは

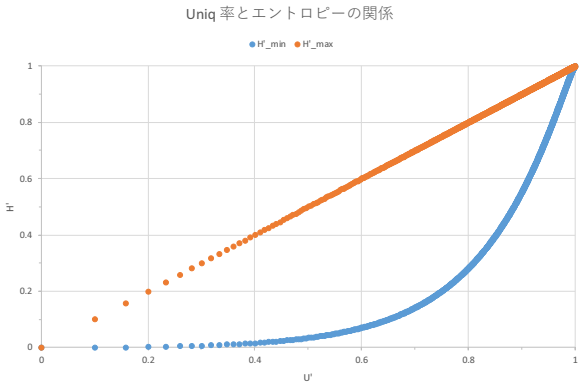

図 1 uniq 率 U'_t の値に対応するエントロピーの上限値 $H'_{t,\max}$ と下限値 $H'_{t,\min}$.

Fig. 1 Maximum $H'_{t,\max}$ and minimum $H'_{t,\min}$ of entropy corresponding to a value of uniq rate U'_t.

uniq 率の実用性に関して一定の目安を与えることが期待できる.

m_t の値に対する正規化されたエントロピー H'_t の上限 $H'_{t,\max}$ の値は，ラグランジュの未定乗数法により求まる．それは，各要素での DNS アクセス回数が n_t/m_t の時であり，

$$\begin{aligned} H'_{t,\max} &= \frac{-\sum_{i\in X_t}^{m_t} \frac{1}{m_t}\log_2\frac{1}{m_t}}{\log_2 n_t} \\ &= \frac{log_2 m_t}{\log_2 n_t} \\ &= \log_{n_t} m_t \\ &= U'_t \end{aligned} \tag{7}$$

となる.

m_t に対するエントロピーの下限 $H'_{t,\min}$ とは一強皆弱状態と考えられる．具体的には，1 種類の DNS アクセス回数数は $n_t-(m_t-1)$ であり，残り m_t-1 種類の DNS アクセス回数数は 1 である．始めに正規化されていないエントロピーでの下限 $H_{t,\min}$ を求め，$H_{t,\min}$ から $H'_{t,\min}$ へ変換すると，

$$\begin{aligned} H_{t,\min} &= \{-\frac{n_t-(m_t-1)}{n_t}\log_2\frac{n_t-(m_t-1)}{n_t}\} \\ &\quad -\sum_{i=1}^{m_t-1}\{\frac{1}{n_t}\log_2\frac{1}{n_t}\} \\ &= \log_2 n_t - \frac{n_t-(m_t-1)}{n_t}\log_2(n_t-(m_t-1)), \\ H'_{t,\min} &= 1-\frac{n_t-(m_t-1)}{n_t}\log_{n_t}(n_t-(m_t-1)) \end{aligned} \tag{8}$$

となる.

式 7 および式 8 を基に, uniq 率 U'_t の各値に対する $H'_{t,\max}$ および $H'_{t,\min}$ の各値を**図 1** に示す．式 8 の中には n_t が含まれることから，本論文では n_t の値として 1,000 を用いる.

図 1 からは，U'_t の値が 0 もしくは 1 に近づくにつれて，$H'_{t,\max}$ の値および $H'_{t,\min}$ の値は近づくことが分かる．このことから，U'_t の実用性は U'_t の値が 0 もしくは 1 に近い時に高まると考えられる.

4. データセットに基づく uniq 率とエントロピー

uniq 率の値とエントロピーの値を比較するため、MIT Lincoln Laboratory および United States Military Academy in West Point で公開されているデータセットを用いて uniq 率の時系列データとエントロピーの時系列データをそれぞれ作成する．作成された uniq 率の時系列データとエントロピーの時系列データを基に，エントロピーの値の変化方向に対応する uniq 率の値の変化方向には一定以上の追随性があるか否かを分析する.

4.1 DARPA2000

MIT Lincoln Laboratory のサイト上に公開されている 2000 DARPA Intrusion Detection Scenario Specific Data Sets のデータセットの中から LLDOS 1.0 inside および LLDOS 2.0.2 inside を入手した [17]．両者は共に DDOS 攻撃によるアクセスデータであり，その違いは攻撃の仕方にある．入手したデータセットは tcpdump 形式のファイルであり各種ネットワークサービスのアクセスログが含まれている．そこで，そのファイルの中から DNS の問合せ部分を抽出した．得られたデータセットの時間長は，それぞれ 190 分間 および 102 分間である．DNS アクセス回数は，それぞれ 71,229 回 および 39,427 回である．その抽出した DNS のアクセスログを基に 1 分間単位で時系列データをそれぞれ作成した．それぞれの時系列データを**図 2** および**図 5** に示す．それら両図からは，それぞれのアクセスの違いを読み取れる．それら両図に示されている各時刻の DNS アクセス回数を基に uniq 率およびエントロピーを計算した．uniq 率およびエントロピーの時系列データを**図 3**, **図 4**, **図 6**, **図 7** にそれぞれ示す.

4.2 CDX2009

United States Military Academy in West Point のサイト上では，演習攻撃を行った際のデータセットが公開されている [18]．公開されているデータセットの中に DNS サーバのアクセスログファイルがある．そこで uniq 率の値とエントロピーの値の比較データとして，そのログファイルも用いた．ログの時間長は 79 時間である．DNS のアクセス回数は 52,852 回である．その抽出した DNS のアクセスログを基に 1 分間単位で時系列データを作成した．その時系列データを**図 8** に示す．その図からは，データの初期と終盤に大量のアクセスがあったことを読み取れる．図 8 に示されている各時刻の DNS アクセス回数を基に uniq 率およびエントロピーを計算した．uniq 率およびエントロピー

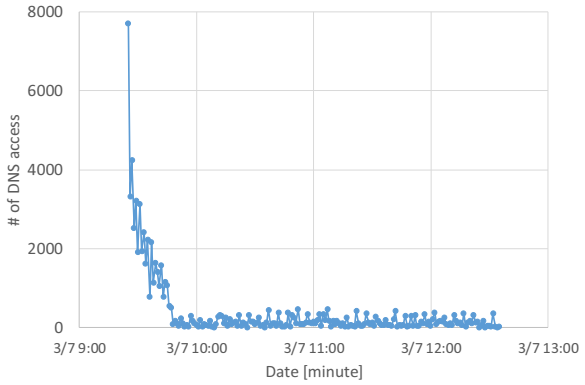

図 2 DARPA 2000 データセットの LLDOS 1.0 inside に基づく DNS アクセス回数の時系列.

Fig. 2 Timeseries of DNS access counts in DARPA 2000 LLDOS 1.0 inside dataset.

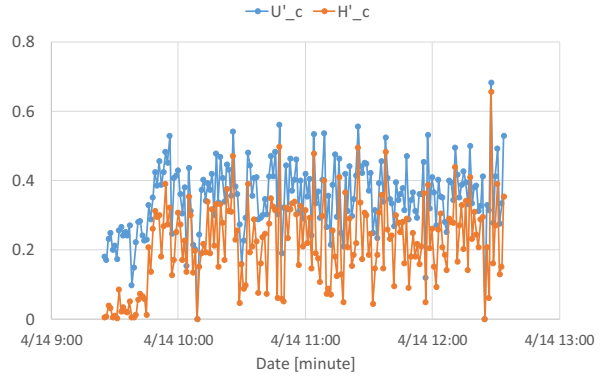

図 3 DARPA 2000 データセットの LLDOS 1.0 inside に基づくクライアントベースの uniq 率およびエントロピーの時系列.

Fig. 3 Timeseries of uniq rate and entropy based on clients in DARPA 2000 LLDOS 1.0 inside dataset.

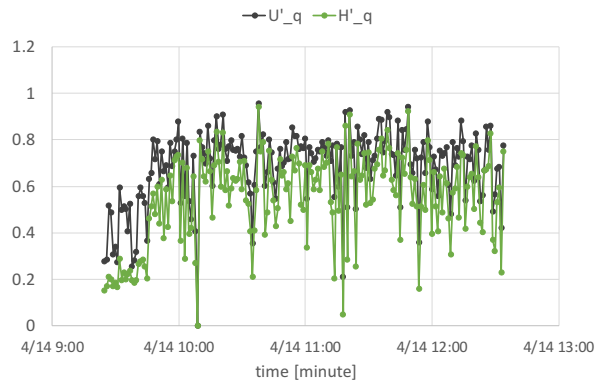

図 4 DARPA 2000 データセットの LLDOS 1.0 inside に基づくクエリベースの uniq 率およびエントロピーの時系列.

Fig. 4 Timeseries of uniq rate and entropy based on queries in DARPA 2000 LLDOS 1.0 inside dataset.

の時系列データを**図 9**, **図 10** にぞれぞれ示す.

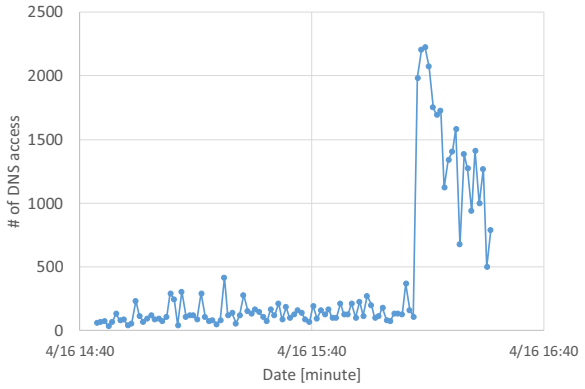

図 5 DARPA 2000 データセットの LLDOS 2.0.2 inside に基づく DNS アクセス回数の時系列.

Fig. 5 Timeseries of DNS access counts of DARPA 2000 LLDOS 2.0.2 inside dataset.

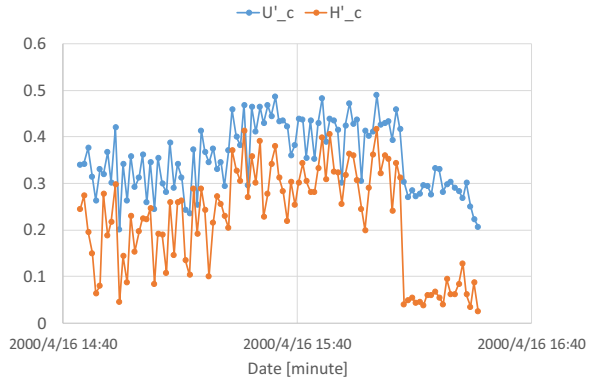

図 6 DARPA 2000 データセットの LLDOS 2.0.2 inside に基づくクライアントベースの uniq 率およびエントロピーの時系列.

Fig. 6 Timeseries of uniq rate and entropy based on clients in DARPA 2000 LLDOS 2.0.2 inside dataset.

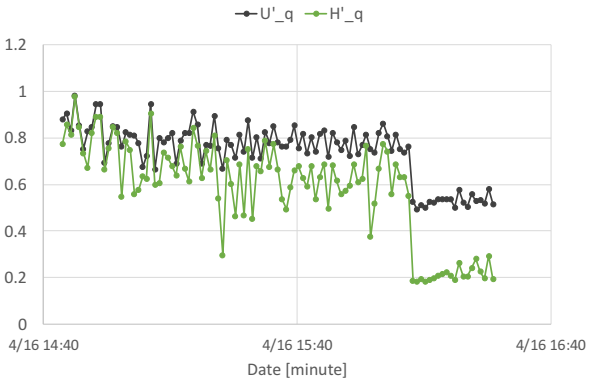

図 7 DARPA 2000 データセットの LLDOS 2.0.2 inside に基づくクエリベースの uniq 率およびエントロピーの時系列.

Fig. 7 Timeseries of uniq rate and entropy based on queries in DARPA 2000 LLDOS 1.0 inside dataset.

4.3 uniq 率のエントロピーに対する追随性

図 3, 図 4, 図 6, 図 7 からは, uniq 率の値の変化方向はエントロピーの値の変化方向に対して一定以上の追随性が

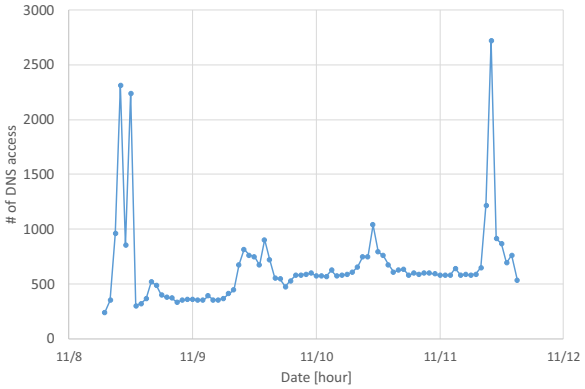

図 8 CDX 2009 データセットに基づく DNS アクセス回数の時系列.

Fig. 8 Timeseries of DNS access counts in CDX 2009 dataset.

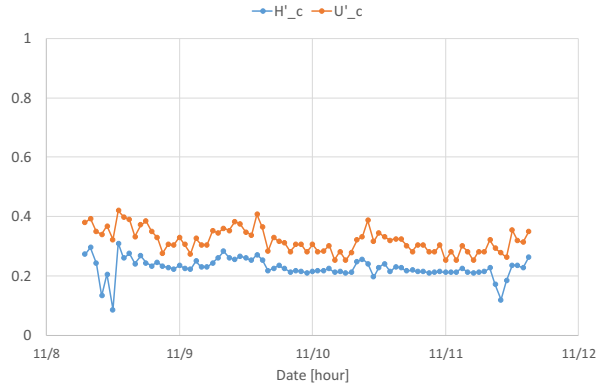

図 9 CDX 2009 データセットに基づくクライアントベースの uniq 率およびエントロピーの時系列.

Fig. 9 Timeseries of uniq rate and entropy based on clients in CDX 2009 dataset.

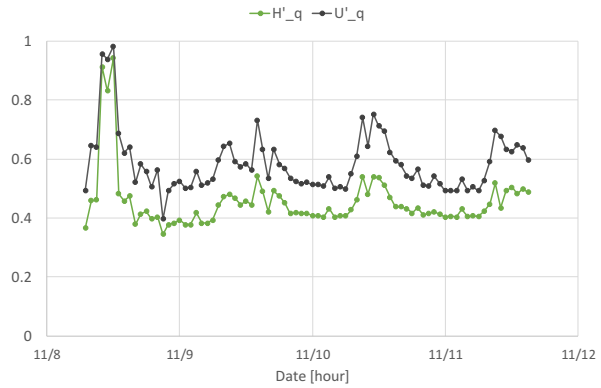

図 10 CDX 2009 データセットに基づくクエリベースの uniq 率およびエントロピーの時系列.

Fig. 10 Timeseries of uniq rate and entropy based on queries in CDX 2009 dataset.

あるように読み取れる．そこで，本論ではそれぞれの値の変化の方向にのみに注目し，uniq 率のエントロピーに対する追随性の定量的な分析を試みた．具体的には以下の手順に従い追随性を分析した．

表 1 uniq 率のエントロピーに対する追随性
Table 1 Traceability of uniq rate for entropy.

Type	DARPA 1.0	DARPA 2.0.2	CDX 2009
Client base	0.6950	0.4496	0.5309
Query base	0.7271	0.5686	0.7314

(1) 時刻 t での uniq 率とエントロピーの各値を 1 つ前の時刻 $t-1$ でのぞれぞれの値との差分を求める．
(2) もし差分値が 0 より大きければ t における変化は $+1$，マイナスであれば変化は -1，0 ならば変化は 0 とする．
(3) 各時刻での変化値 -1, 0, $+1$ を基に時系列データを作成し，uniq 率側とエントロピー側との間で相関係数を求める．
(4) 以上の手順をクライアントベースおよびクエリベースのそれぞれで行う．

分析結果を**表 1** に示す．表 1 からは，uniq 率はエントロピーとの間に一定以上の追随性を読み取れる．

5. まとめ

本論では，外れ値検知の 1 手段として DNS サーバへのアクセスログから求められた uniq 率とエントロピーとの比較検討を行った．uniq 率は，単位時間当りの DNS サーバへのアクセス数に対する，クライアント数もしくはクエリの種類数の割合である．uniq 率は，エントロピーに基づく外れ値検知の前処理として，エントロピーよりも少ない計算量で外れ値検知を行うことが期待できる．

uniq 率は，その定義上，算出した値に対応するエントロピーの値に幅を有する．今回，その幅の程度を理論的に考察し，uniq 率の取り得る値の両端ではその幅は小さくなっていくことが分かった．

uniq 率とエントロピーとの関係をさらに調査するため，CDX 2009 データセットおよび DARPA 2000 データセットの 2 つに含まれる DNS アクセスデータから uniq 率とエントロピーの時系列データをそれぞれ作成した．両時系列データの値の変化に対する追随性を調べるため，両者の値の変動方向を基に相関係数を算出したところ，両者の間には一定以上の相関がある事が分かった．

参考文献

[1] Feinstein, L., Schnackenberg, D., Balupari, R. and Kindred, D.: Statistical approaches to DDoS attack detection and response, *DARPA Information Survivability Conference and Exposition, 2003. Proceedings*, IEEE, pp. 303–314 (2003).
[2] Hodge, V. J. and Austin, J.: A Survey of Outlier Detection Methodologies, *Artificial Intelligence Review*, Vol. 22, pp. 85–126 (2004).
[3] Celenk, M., Conley, T., Willis, J. and Graham, J.: Anomaly detection and visualization using Fisher Discriminant clustering of network entropy, *Digital Infor-*

mation Management, 2008. ICDIM 2008. Third International Conference on, IEEE, pp. 13–16 (2008).
[4] Lee, K., Kim, J., Kwon, K. H., Han, Y. and Kim, S.: DDoS attack detection method using cluster analysis, *Expert Systems with Applications*, Vol. 34, pp. 1659–1665 (online), DOI: 10.1016/j.eswa.2007.01.040 (2008).
[5] Lu, K., Wu, D., Fan, J., Todorovic, S. and Nucci, A.: Robust and efficient detection of DDoS attacks for large-scale internet, *Computer Networks*, Vol. 51, pp. 5036–5056 (online), DOI: 10.1016/j.comnet.2007.08.008 (2007).
[6] Xiao, B., Chen, W. and He, Y.: An autonomous defense against SYN flooding attacks: Detect and throttle attacks at the victim side independently, *Journal of Parallel and Distributed Computing*, Vol. 68, pp. 456–470 (online), DOI: 10.1016/j.jpdc.2007.06.013 (2008).
[7] 山西健司：データマイニングによる異常検知，共立出版 (2009).
[8] 松原義継，武藏泰雄：DNS アクセスの uniq 率に基づく外れ値検知の試み，第 8 回インターネットと運用技術シンポジウム (IOTS2015)，Vol. 8, pp. 1–5（オンライン），入手先 ⟨http://id.nii.ac.jp/1001/00145973/⟩ (2015).
[9] Shomura, Y., Watanabe, Y. and Yoshida, K.: Analyzing the number of varieties in frequently found flows, *IEICE Transactions on Communications*, Vol. E91-B, No. 6, pp. 1896–1905 (online), DOI: 10.1093/ietcom/e91-b.6.1896 (2008). 被引用数 8.
[10] 吉田健一，三田村健史：ネットワークデータのオンライン異なり数解析 (＜特集＞データ中心科学)，人工知能:人工知能学会誌，Vol. 30, No. 2, pp. 230–237（オンライン），入手先 ⟨http://ci.nii.ac.jp/naid/110009913059/⟩ (2015).
[11] 三田村健史，吉田健一：DNS クエリデータに基づくコンテンツへの関心度分析 (＜特集＞社会基盤としてのインターネットアーキテクチャ論文)，電子情報通信学会論文誌. B, 通信， Vol. 93, No. 10, pp. 1368–1377（オンライン），入手先 ⟨http://ci.nii.ac.jp/naid/110007730039/⟩ (2010).
[12] 吉田健一：ネットワークデータの異なり数計測とその応用，オープンデータとセキュリティ，サイエンティフィック・システム研究会 (2015).
[13] Takeda, Y., Musashi, Y. and Moriyama, K. S. T.: DNS ANY Request Cannon Activity in DNS Query Packet Traffic, *International Journal of Intelligent Engineering and Systems*, Vol. 7, No. 1, pp. 8–16 (2014).
[14] Musashi, Y., Takeda, Y., Shibata, N., Kubota, S. and Sugitani, K.: A Statistical Study of ANY Resource Record Based DNS Query Request Packet Traffic, *Information*, Vol. 16, No. 12(B), pp. 8901–8908 (2013).
[15] Takemori, K., Kong, W. J., na Romaña, D. A. L., Kubota, S., Sugitani, K. and Musashi, Y.: Entropy Study on A Resource Record Query Traffic from the Campus Network, *IPSJ SIG Technical Reports, Internet Operation and Technology 4th (IOT4)*, Vol. 2009, No. 21, pp. 101–106 (2009).
[16] na Romaña, D. A. L. and Musashi, Y.: Entropy Based Analysis of DNS Query Traffic in the Campus Network, *Proceedings of The 4th International Conference on Cybernetics and Information Technologies, System and Applications (CITSA2007)*, Vol. 6, No. 5, pp. 162–164 (2007).
[17] Laboratory, M. L.: 2000 DARPA Intrusion Detection Scenario Specific Data Sets. http://www.ll.mit.edu/ideval/data/2000data.html.
[18] in West Point, T. U. S. M. A.: CDX 2009 Data Sets. http://www.usma.edu/crc/sitepages/datasets.aspx.