Uso de padrões ordinais na caracterização e análise de ataques de botnets em Internet das Coisas (IoT)

Eduarda T. C. Chagas
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brasil
eduarda.chagas@dcc.ufmg.br

João B. Borges
Departamento de Computação e Tecnologia
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
joao.borges@ufrn.br

Heitor S. Ramos
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brasil
ramosh@dcc.ufmg.br

Abstract

The main objective of this tutorial is to present the fundamentals of temporal data analysis using ordinal patterns and descriptors from Information Theory, covering the tools and steps necessary for developing applications and services to detect botnets in the Internet of Things (IoT) scenarios. Thus, we investigated and presented the solutions proposed in the literature for the following questions: (i) What are the main advantages of the Bandt-Pompe methodology in the temporal data analysis process? (ii) How can we use Information Theory descriptors and ordinal patterns in data characterization activities? (iii) What are the main research problems? (iv) What are the main characteristics of the methodology that enable the development of applications in the context of IoT and botnet detection?

Keywords: Ordinal patterns, Botnet detection, Internet of Things, Information Theory

1 Introdução

Nos últimos anos, temos visto um crescimento significativo no número de aplicações inteligentes envolvendo análise, mineração e classificação de dados. Consequentemente, causando um aumento na diversidade e no volume de informações utilizadas. Com isso, aumentou-se também o nível de complexidade das investigações, a interdisciplinaridade e o número de recursos necessários para a realização de tais atividades. Assim, o estudo de abordagens simples, computacionalmente baratas e independentes do tipo de dados para extração e caracterização de padrões tornou-se fundamental.

Dados provenientes de séries temporais carregam informações valiosas sobre o sistema que produz os dados e sua análise, tradicionalmente, é baseada em uma das duas grandes frentes: nos domínios do tempo (natural) e no domínio transformado (por exemplo, frequência e *wavelet*) [7]. No entanto, as principais técnicas de ambas abordagens utilizam diretamente os dados resultantes do processo observacional, que são sensíveis a ruídos e efeitos de contaminação. Em contrapartida, [2] propuseram a transformação em padrões ordinais, uma abordagem não paramétrica para análise de séries temporais que consiste em registrar a ordem relativa das observações de pequenos subconjuntos da série temporal. O uso de padrões ordinais obtidos quando combinados com descritores da teoria da informação mostraram ter um alto poder de caracterização do processo subjacente à dinâmica dos dados, sendo aplicados em caracterização de séries temporais nas mais diversas áreas do conhecimento.

Por outro lado, um ataque amplamente realizado em dispositivos presentes na Internet das Coisas (*Internet of Things* – IoT), é sua infecção por bots, que possui como principal ameaça à segurança a possibilidade de ataques orquestrados em grande escala, como entrega de spam e ataques distribuídos de negação de serviço (DDoS) [6]. Sob esse contexto, ao longo deste tutorial apresentaremos uma introdução à análise de séries temporais com o uso de padrões ordinais extraídos através da simbolização de Bandt-Pompe, o estado-da-arte e uma visão geral dos trabalhos existentes sobre a metodologia e sua aplicação em diferentes cenários de IoT, assim como os desafios e as perspectivas futuras para o avanço da área.

1.1 Transformação ordinal de Bandt-Pompe

Seja $\mathcal{X} \equiv \{x_t\}_{t=1}^{T}$ uma série temporal de comprimento T, podemos descrever a simbolização de Bandt-Pompe como o processo de obtenção do conjunto de padrões ordinais presentes ao longo de uma dada sequência. Para capturar tais informações de ordem, os seguintes passos devem ser realizados:

- **Particionamento**: A série inicialmente será particionada em subconjuntos de dimensão D, com τ intervalos entre os pontos da série, possuindo assim a seguinte estrutura:
$$\mathbf{X}_t^{(D,\tau)} = (x_t, x_{t+\tau}, \ldots, x_{t+(D-1)\tau}),$$
para $t = 1, 2, \ldots, N$, onde $N = T - (D-1)\tau$.
- **Permutação**: Os padrões ordinais serão obtidos através da permutação dos índices de cada partição, de

In: Tutoriais do WebMedia, Curitiba, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2022.
© 2022 SBC – Sociedade Brasileira de Computação.
ISSN 2596-1683

forma que estes elementos sejam ordenados de forma crescente.

Tradicionalmente, a partir do conjunto de padrões ordinais obtidos de uma determinada sequência podemos seguir com dois tipos de análise: a distribuição marginal dos padrões (histograma de frequências) e suas transições [4].

Ao optar por essa nova representação dos dados, a distribuição resultante torna-se menos sensível a *outliers* e, como não depende de nenhum modelo, pode ser aplicada a diversas situações. Apesar de sua simplicidade, este método é robusto ao ruído e apresenta bons resultados na avaliação da aleatoriedade de uma sequência, caracterização e classificação de sinais.

1.1.1 Histograma de frequências. A abordagem clássica para calcular a distribuição de probabilidade de padrões ordinais é através do histograma de frequência. Seja Π a sequência de símbolos obtida por uma dada série $\mathbf{X}_t^{(D,\tau)}$. A distribuição de probabilidade Bandt-Pompe será dada pela frequência relativa dos $D!$ possíveis padrões $\{\widetilde{\pi}_t^D\}_{t=1}^{D!}$:

$$p(\widetilde{\pi}_t^D) = \frac{\#\left\{\mathbf{X}_t^{(D,\tau)} \text{ é do tipo } \widetilde{\pi}_t^D\right\}}{T-(D-1)\tau}, \tag{1}$$

onde $t \in \{1, \ldots, T-(D-1)\tau\}$. Essas probabilidades atendem às condições $p(\widetilde{\pi}_t^D) \geq 0$ e $\sum_{i=1}^{D!} p(\widetilde{\pi}_t^D) = 1$, e são invariantes a transformações monotônicas dos valores das séries temporais. Por exemplo, a presença de ruído multiplicativo α em $\mathcal{X}$ não altera os resultados dos padrões produzidos.

1.1.2 Grafos de Transição. Alternativamente, também podemos analisar o conjunto de padrões ordinais como um grafo orientado $G = (V, E)$ que represente as transições entre dois padrões ordinais consecutivos ao longo do tempo t. Nesta nova representação, cada vértice corresponderá a um símbolo e as arestas às transições entre estes: $V = \{v_{\widetilde{\pi}_t^D}\}$, e $E = \{(v_{\widetilde{\pi}_t^D}, v_{\widetilde{\pi}_{t+1}^D}) : v_{\widetilde{\pi}_t^D}, v_{\widetilde{\pi}_{t+1}^D} \in V\}$ [5]. Os pesos $W = \{w_{v_{\widetilde{\pi}_i^D}, v_{\widetilde{\pi}_j^D}} : v_{\widetilde{\pi}_i^D}, v_{\widetilde{\pi}_j^D} \in V\}$ atribuídos a cada aresta descreve a probabilidade de existência de transições entre os padrões $(v_{\widetilde{\pi}_i^D}, v_{\widetilde{\pi}_j^D})$ ao longo da sequência, ou seja, os pesos são calculados como a frequência relativa de cada transição:

$$w_{v_{\widetilde{\pi}_i^D}, v_{\widetilde{\pi}_j^D}} = \frac{|\Pi_{\widetilde{\pi}_i^D, \widetilde{\pi}_j^D}|}{T-(D-1)\tau-1}, \tag{2}$$

onde $|\Pi_{\widetilde{\pi}_i^D, \widetilde{\pi}_j^D}|$ é o número de transições do padrão $\widetilde{\pi}_i^D$ para o padrão $\widetilde{\pi}_j^D$, $\sum_{v_{\widetilde{\pi}_i^D}, v_{\widetilde{\pi}_j^D}} w_{v_{\widetilde{\pi}_i^D}, v_{\widetilde{\pi}_j^D}} = 1$, e o denominador é o número de transições presentes na série de comprimento $T-(D-1)\tau$.

1.2 Plano Entropia-Complexidade

Um dos campos de estudo nesta área é a aplicação de descritores da Teoria da Informação. A Teoria da Informação desenvolvida por Claude Shannon surgiu como um ramo interdisciplinar, produzindo inúmeros resultados tanto do ponto de vista teórico quanto nas aplicações na extração de informação em sinais. Atualmente, consiste em uma poderosa ferramenta para a quantificação dos diferentes níveis de ordem e complexidade presentes nos processos que geram os dados.

O uso de padrões ordinais quando combinados com o uso de descritores da Teoria da Informação mostraram ter um alto poder de discriminação do processo subjacente à dinâmica dos dados. Dentre os descritores, dois deles por apresentarem definições complementares têm recebido grande destaque na literatura: a Entropia de Shannon e a Complexidade Estatística. Por exemplo, seja $P = \{p_{\widetilde{\pi}_1^D}, p_{\widetilde{\pi}_2^D}, \ldots, p_{\widetilde{\pi}_{D!}^D}\} = \{p_1, \ldots, p_{D!}\}$ a distribuição de probabilidade obtida a partir do histograma de frequências de padrões ordinais. A Entropia de Permutação de Shannon mensurará o grau de desordem e impreditibilidade de um sistema, sendo calculada utilizando a seguinte fórmula:

$$H(P) = -\sum_{i=1}^{D!} p_{\widetilde{\pi}_i^D} \log p_{\widetilde{\pi}_i^D}. \tag{3}$$

No entanto, na literatura opta-se por utilizar a entropia de Shannon normalizada definida por [13]:

$$H_S(\mathbb{P}) = \frac{H(\mathbb{P})}{H_{max}} = -\frac{1}{\log D!} \sum_{i=1}^{D!} p_{\widetilde{\pi}_i^D} \log p_{\widetilde{\pi}_i^D}, \tag{4}$$

onde $0 \leq H_S(\mathbb{P}) \leq 1$.

Já a complexidade estatística, será dada pelo produto da Entropia de Permutação e uma medida de "desequilíbrio", que consiste na distância para uma distribuição de equilíbrio, que geralmente obedece a lei uniforme. A complexidade estatística quantifica, além da aleatoriedade, a presença de estruturas correlacionais entre os padrões. Desta forma, diferentes graus de estrutura podem ser quantificados, refletindo propriedades reveladas pela distribuição de probabilidade do processo subjacente e é dada por:

$$C_{JS}(P, U) = H_S(P) \cdot Q_{JS}(P, U), \tag{5}$$

onde $U = (\frac{1}{D!}, \ldots, \frac{1}{D!})$ corresponde a distribuição uniforme.

A medida de "disequilibrium" Q_{JS} é calculada como:

$$Q_{JS}(P, U) = Q_0 \cdot \left\{ H\left[\frac{P+U}{2}\right] - \frac{H[P+U]}{2} \right\}, \tag{6}$$

onde a constante de normalização é dada por:

$$Q_0 = -2\left\{\left(\frac{D!+1}{D!}\right)\ln(D!+1) - 2\ln(2D!) + \ln(D!)\right\}^{-1}, \tag{7}$$

de modo que $0 \leq Q_0 \leq 1$.

O conjunto de todos os pares $(H_S(P), C_{JS}(P, U))$ para qualquer série temporal descrita por padrões de comprimento D está em um subconjunto compacto de R^2: o plano Entropia-Complexidade (ou plano Complexidade-Entropia, $H \times C$). Através desta ferramenta é possível descobrir a natureza da

série apenas verificando sua região de localização no plano, pois seus valores associados ajudam a determinar se ela corresponde a uma sequência caótica, determinística ou estocástica.

1.3 Desafios e Oportunidades de Pesquisa

A análise de padrões ordinais foi proposta inicialmente em meados de 2002 e desde então houve uma evolução da técnica e sua capacidade de aplicação nas mais diversas áreas do conhecimento. Como podemos ver no esquema da Figura 1, muitos resultados foram obtidos com a aplicação direta da metodologia em atividades de caracterização de dados temporais. No entanto, ao longo dos anos, a mudança de foco ocorreu no sentido de propor soluções para lacunas no artigo seminal. Entre algumas das limitações investigadas estão a técnica aplicada a dados com mais de uma dimensão, a adição de informações de amplitude do sinal e a imputação de elementos repetidos junto com a série temporal. Por outro lado, também pudemos observar certa tendência na automação, com o surgimento de propostas de classificação, previsão e seleção de hiperparâmetros. Como resultado, o uso de algoritmos de aprendizado de máquina está se tornando mais frequente.

Sob esse contexto, a seguir discutiremos alguns problemas de pesquisa que são de grande interesse da comunidade de análise de padrões ordinais utilizando a simbolização de Bandt-Pompe. Não temos como objetivo exaurir todos os problemas de pesquisa em aberto na área, desse modo, para uma discussão mais completa, recomendamos o artigo de [8].

1.3.1 Deep Learning. O uso de inteligência artificial com simbolização de padrões ordinais e descritores de Teoria da Informação ainda é incipiente, destacando-se no uso de algoritmos de aprendizado de máquina (como k-vizinhos mais próximos, máquina de vetores de suporte, algoritmos baseados em árvore, etc.) em processos de caracterização de dados. Dentre as metodologias aplicadas, destacam-se aquelas obtidas a partir do plano complexidade-entropia. Assim, podemos ver que o uso de técnicas de deep learning para aumentar o poder discriminativo de padrões ordinais ainda é um problema em aberto e ainda pouco investigado. Portanto, um campo frutífero de pesquisa.

1.3.2 Propriedades estatísticas do plano $H \times C$. Embora os limites de $H \times C$ estejam bem definidos, uma caracterização completa de sua topologia intrínseca é um problema em aberto, devido às restrições impostas pelo seu espaço curvilíneo. O desconhecimento da distribuição conjunta dos pontos obtidos por este plano, devido à correlação existente entre suas variáveis, inviabiliza os estudos de estatísticas de teste para séries temporais típicas neste espaço de caracterização. No entanto, com o conhecimento da variabilidade esperada de tais pontos, de acordo com a dinâmica subjacente, podemos testar hipóteses para uma grande variedade de modelos. Resultados nessa direção podem ser encontrados na literatura. [12] mostrou que o plano Entropia-Complexidade é um bom indicador dos resultados dos testes Diehard para geradores de números pseudo-aleatórios. [10] avaliou formas de melhorar sequências pseudo-aleatórias para sua representação neste plano. Nesse contexto, um problema em aberto presente na caracterização de sequências usando o plano $H \times C$ é a ausência de uma distância métrica representativa, o que dificulta a construção de regiões de confiança.

1.3.3 Padrões Ordinais Ponderados. Embora o uso da simbolização de Bandt-Pompe tenha várias vantagens em relação a outros algoritmos de extração de características, ela possui duas lacunas principais em sua definição original que foram abordadas por meio de métodos de ponderação de padrões ordinais: (i) A ambiguidade ordinal presente quando temos valores iguais na mesma subsequência, e (ii) A falta de informações relacionadas às diferenças na amplitude da amostra, ou seja, o valor médio das amplitudes e as diferenças entre amostras vizinhas não são considerados pela metodologia original. No entanto, ainda existe uma lacuna de estudos que busquem pesos ótimos capazes de aumentar o poder de discriminação da técnica no plano de Entropia-complexidade por meio de métodos de otimização.

1.4 Aplicações

A metodologia Bandt-Pompe e suas variantes têm sido utilizadas com sucesso na análise de diversos tipos de dinâmicas, recebendo até agora mais de 2160 citações advindas de 684 periódicos distintos, de acordo com o Web of Science[1]. Encontramos trabalhos utilizando essa abordagem em diversas áreas do conhecimento científico, dentre elas, análise de dados sob o contexto de IoT, como por exemplo: o mapeamento da dinâmica do tráfego de rede [11]; a caracterização de eletrodomésticos através do seu consumo de energia [1]; proposta de amostragem de dados para aplicações de cidades inteligentes baseado em infraestrutura de rede de sensores [15] e estudos sobre velocidades de veículos e identificação de meios de transporte com base em informações de GPS [9, 16].

1.5 Estudo de caso - Hands On

Para colocar em prática os conhecimentos adquiridos, finalizaremos com uma aplicação no contexto de detecção de ataques de botnet em IoT. Utilizaremos o conjunto de dados públicos N-BaIoT que consistem no tráfego de rede coletado de nove dispositivos IoT durante operações regulares (benigno) e sob ataques de botnet (anomalias) [14]. O tráfego de rede anômalo foi gerado infectando cada dispositivo com as famílias de botnet Mirai e Bashlite [3]. Nosso principal objetivo será identificar anomalias na dinâmica temporal de tais dispositivos IoT. A modelagem do problema será realizada utilizando-se o pacote NATS[2].

[1] Verificado em 23 de outubro de 2021
[2] https://github.com/EduardaChagas/NATS

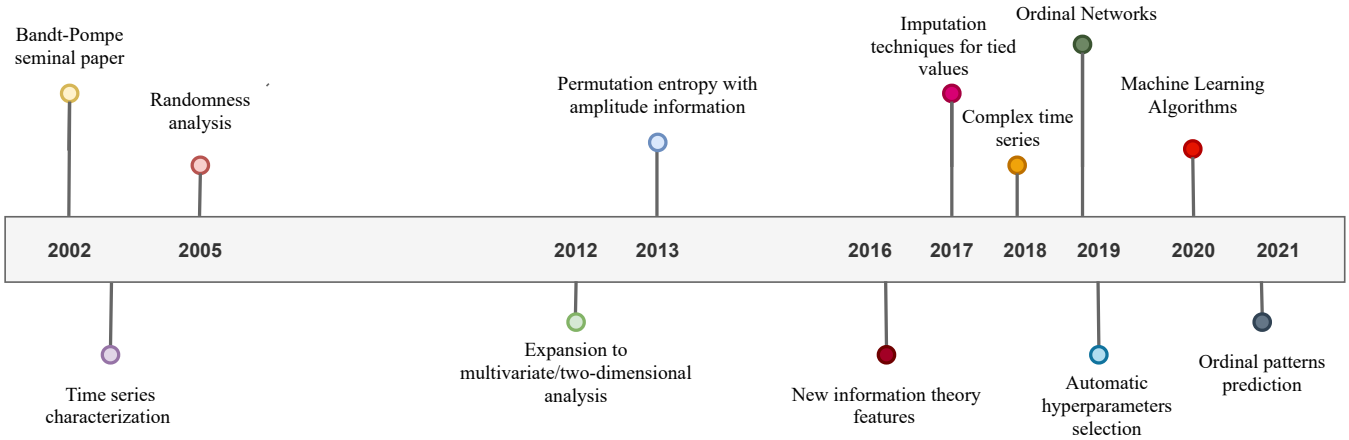

Figura 1. Visão geral da evolução temporal dos temas investigados em análise de padrões ordinais.

1.5.1 Ataques de botnet em IoT. Devido aos seus recursos computacionais limitados, dispositivos IoT apresentam uma maior vulnerabilidade a ataques de segurança, incluindo a sua infecção por bots. Um bot consiste em um software malicioso para controlar remotamente dispositivos por um invasor, o botmaster, sendo a rede de dispositivos comprometidos por bots denominada botnet. As botnets são ameaças com um alto grau de severidade, pois comumente acionam ataques de negação de serviço, que são amplificados por um grande número de dispositivos IoT [6]. Logo, nesta aplicação descrevemos como podemos utilizar transformações de padrões ordinais e descritores oriundos da Teoria da Informação, obtidos a partir de dados do número de pacotes transmitidos por dispositivos ao longo do tempo, para identificar botnets em dispositivos IoT.

1.5.2 Caracterização de botnets. O NATS[3] (Non-parametric Analysis Time Series) é uma ferramenta portátil, rápida e de boa qualidade numérica que possibilita a análise interativa e exploratória de dados de uma série temporal por meio de técnicas da Teoria da Informação. Com ele, o usuário dispõe de um conjunto de técnicas de análise presentes na literatura para processar e examinar seus dados de forma eficiente e com um tempo mínimo de aprendizado. Além de fornecer um pacote de funcionalidades em R (utilizado neste presente tutorial), a ferramenta é extensível e apresenta um sistema interativo web. Todo o código fonte e dados utilizados se encontram disponíveis em nosso repositório presente no Github[4].

Com os resultados obtidos, observamos que ao utilizarmos a distribuição de padrões ordinais obtida pelo histograma da frequência e calcularmos seus respectivos pontos no plano Entropia-Complexidade, conseguimos discriminar a dinâmica temporal de algumas operações de ataques regulares. No entanto, para uma melhor distinção, aplicamos uma análise multiescala em conjunto com as probabilidades de auto-transição obtidas ao seguir a abordagem de análise transacional. Com isso, observamos que à medida que τ aumenta, maior a discriminação entre as operações de ataque regulares e outras do dispositivo. Análise multiescala de padrões ordinais consiste na utilização de diferentes tempos de atraso (*delay*) no processo de simbolização e extração de características. Assim, conseguimos analisar a dependência e correlação temporal entre os elementos de uma dada sequência, auxiliando na correta caracterização da dinâmica geradora subjacente.

[3]https://github.com/EduardaChagas/NATS

[4]https://github.com/EduardaChagas/ordinal_pattern_botnet

1.6 Considerações Finais

Este tutorial introduziu a metodologia de simbolização de Bandt-Pompe no contexto de Internet das Coisas e análise de ataques de botnets em Internet das coisas. Inicialmente, apresentamos uma breve introdução a padrões ordinais e os descritores de Teoria da Informação mais utilizados na literatura para caracterização de séries temporais. Em seguida, foram apresentados alguns dos principais desafios de pesquisa e listadas algumas das mais recentes aplicações da área no contexto de IoT. Finalmente, introduzimos a problemática de detecção de botnets e propomos um estudo de caso com enfoque na caracterização das operações de ataque aos dispositivos.

Referências

[1] A. L. L. Aquino, H. S. Ramos, A. C. Frery, L. P. Viana, T. S. G. Cavalcante, and O. A. Rosso. 2017. Characterization of Electric Load with Information Theory Quantifiers. *Physica A* 465 (2017), 277–284. https://doi.org/10.1016/j.physa.2016.08.017

[2] C. Bandt and B. Pompe. 2002. Permutation Entropy: A Natural Complexity Measure for Time Series. *Physical Review Letters* 88 (2002), 174102–1–174102–4. Issue 17. https://doi.org/10.1103/PhysRevLett.88.174102

[3] Elisa Bertino and Nayeem Islam. 2017. Botnets and internet of things security. *Computer* 50, 2 (2017), 76–79.

[4] J. Borges, H. Ramos, R. Mini, O. A. Rosso, A. C. Frery, and A. A. F. Loureiro. 2019. Learning and Distinguishing Time Series Dynamics via Ordinal Patterns Transition Graphs. *Appl. Math. Comput.* 362 (2019), UNSP 124554. https://doi.org/10.1016/j.amc.2019.06.068

[5] J. Borges, H. Ramos, R. Mini, O. A. Rosso, A. C. Frery, and A. A. F. Loureiro. 2019. Learning and Distinguishing Time Series Dynamics via Ordinal Patterns Transition Graphs. *Appl. Math. Comput.* 362 (2019), UNSP 124554. https://doi.org/10.1016/j.amc.2019.06.068

[6] Joao Batista Borges, Joao PS Medeiros, Luiz PA Barbosa, Heitor S Ramos, and Antonio A Loureiro. 2022. IoT Botnet Detection based on Anomalies of Multiscale Time Series Dynamics. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[7] P. J. Brockwell and R. A. Davis. 1991. *Time Series: Theory and Methods* (2 ed.). Springer-Verlag, Berlin.

[8] Inmaculada Leyva Callejas, Johann Martinez, Cristina Masoller, Osvaldo A Rosso, and Massimiliano Zanin. 2022. 20 years of ordinal patterns: Perspectives and challenges. *Europhysics Letters* (2022).

[9] Isadora Cardoso-Pereira, João B Borges, Pedro H Barros, Antonio F Loureiro, Osvaldo A Rosso, and Heitor S Ramos. 2022. Leveraging the self-transition probability of ordinal patterns transition network for transportation mode identification based on GPS data. *Nonlinear Dynamics* 107, 1 (2022), 889–908.

[10] L. De Micco, C. M. González, H. A. Larrondo, M. T. Martin, A. Plastino, and O. A. Rosso. 2008. Randomizing nonlinear maps via symbolic dynamics. *Physica A: Statistical Mechanics and its Applications* 387, 14 (2008), 3373–3383. https://doi.org/10.1016/j.physa.2008.02.037

[11] Cristopher GS Freitas, Osvaldo A Rosso, and Andre LL Aquino. 2020. Mapping Network Traffic Dynamics in the Complexity-Entropy Plane. In *2020 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 1–6.

[12] H. A. Larrondo, M. T. Martín, C. M. González, A. Plastino, and O. A. Rosso. 2006. Random number generators and causality. *Physics Letters A* 352, 4–5 (2006), 421–425. https://doi.org/10.1016/j.physleta.2005.12.009

[13] M. T. Martin, A. Plastino, and O. A. Rosso. 2006. Generalized statistical complexity measures: Geometrical and analytical properties. *Physica A: Statistical Mechanics and its Applications* 369, 2 (2006), 439–462. https://doi.org/10.1016/j.physa.2005.11.053

[14] Yair Meidan, Michael Bohadana, Yael Mathov, Yisroel Mirsky, Asaf Shabtai, Dominik Breitenbacher, and Yuval Elovici. 2018. N-baiot—network-based detection of iot botnet attacks using deep autoencoders. *IEEE Pervasive Computing* 17, 3 (2018), 12–22.

[15] Givanildo L Nascimento, Cristopher GS Freitas, Osvaldo A Rosso, and Andre LL Aquino. 2021. Data Sampling Algorithm Based on Complexity-Entropy Plane for Smart Sensing Applications. *IEEE Sensors Journal* 21, 22 (2021), 25831–25842.

[16] Mauricio J Silva, Tamer SG Cavalcante, Osvaldo A Rosso, Joel JPC Rodrigues, Ricardo AR Oliveira, and Andre LL Aquino. 2019. Study about vehicles velocities using time causal information theory quantifiers. *Ad Hoc Networks* 89 (2019), 22–34.