Information Sciences 408 (2017) 1–22

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Mining repeating pattern in packet arrivals: Metrics, models, and applications

Jianfeng Li[a], Xiaobo Ma[a,*], Junjie Zhang[b], Jing Tao[a], Pinghui Wang[a], Xiaohong Guan[a]

[a] MOE KLINNS Lab, Xi'an Jiaotong University, Xi'an, China
[b] Department of Computer Science and Engineering, Wright State University, Dayton, USA

ARTICLE INFO

Article history:
Received 3 August 2016
Revised 13 January 2017
Accepted 19 April 2017
Available online 25 April 2017

Keywords:
Repeating pattern
Temporal structure
Hierarchical clustering
Traffic modeling

ABSTRACT

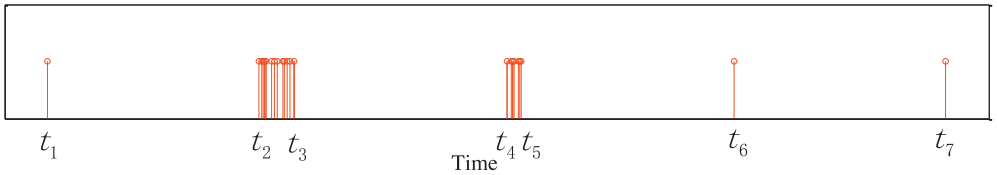
A substantial portion of the network traffic can be attributed to autonomous network applications that experience repeating networking patterns. This observation is further signified by the emergence of the Internet of Things (IoT) era that features an enormous number of networked, autonomous sensors. Identifying and characterizing repeating patterns therefore become a critical means to Internet measurement and traffic engineering. In this paper, we propose a novel method that can effectively identify and characterize timing-based repeating patterns from network traffic by overcoming three significant practical challenges, including i) time-scale sensitive, ii) transience, and iii) being interleaved by noises. Our method features a novel metric, namely unpredictability index (UPI), to capture repeating patterns by quantifying the predictability of packet arrivals' temporal structure from the perspective of hierarchical clustering. An online approach is further developed to incrementally compute UPI upon observing a single packet. Extensive experiments based on synthetic and real-world data have demonstrated that our method can effectively conduct repeating pattern mining.

1. Introduction

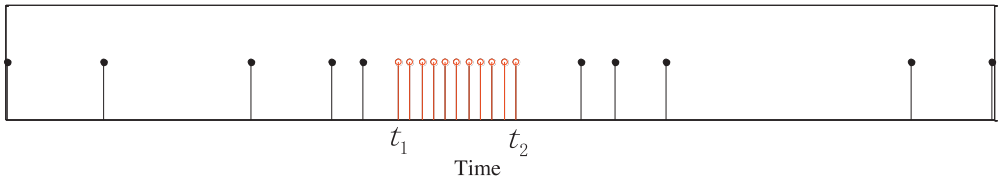
A substantial portion of the network traffic can be attributed to applications that engage in autonomous network activities, where examples of such applications include email clients, antivirus software, and even botnets [41]. This trend is further signified by the emergence of the Internet of Things (IoT) era that features an enormous number of networked, autonomous sensors. It therefore becomes highly demanded to identify and characterize network activities associated with autonomous applications in a variety of areas such as Internet measurement [8], traffic engineering [31], and intrusion detection [22]. A typical feature of autonomous applications is that their network activities usually experience *repeating patterns* (i.e., certain patterns recur in network traffic). For example, an email client periodically contacts a server to retrieve new emails; the antivirus software pings its home server daily for possible updates; a bot regularly sends requests to a command & control (C&C) server to retrieve commands issued by attackers.

* Corresponding author.
E-mail addresses: jfli.xjtu@gmail.com (J. Li), xma.cs@xjtu.edu.cn (X. Ma), junjie.zhang@wright.edu (J. Zhang), jtao@xjtu.edu.cn (J. Tao), phwang@xjtu.edu.cn (P. Wang), xhguan@xjtu.edu.cn (X. Guan).

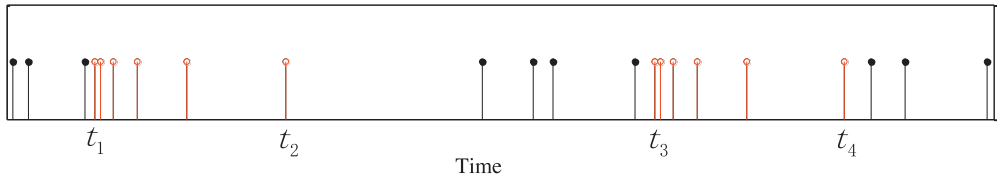
http://dx.doi.org/10.1016/j.ins.2017.04.033
0020-0255/



(a) An example of time-scale sensitive repeating pattern.



(b) An example of transient repeating pattern.



(c) An example of repeating pattern interleaved by noises.

Fig. 1. Illustrations of three temporal properties of repeating pattern. Each bar represents a packet arrival, where the red ones with open circles denote those involved in repeating pattern. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A pattern can characterize network traffic from different aspects such as packet content, packet length, and packet arrival time. Among these, timing-based repeating patterns are particularly useful since they are independent to packet payload and thereby immune to common networking techniques such as fragmentation and encryption that commonly obfuscate both the packet length and content. However, despite the promise, mining timing-based repeating patterns from packet arrivals in real networks is faced with great challenges:

1) Time-Scale Sensitive. The identification of repeating patterns is sensitive to the time scales used. For example, an email client will issue a request to its email server every 15 min to check new emails; if a new email is available, the client will download the email, leading to a burst of packets exchanged between the client and the server. It is worth noting packets within each burst usually do not experience any repeating patterns since their time intervals are jointly influenced by both TCP/IP stack implementation of end points, network congestions, and routing dynamics. Fig. 1(a) visualizes this example. Specifically, packets in t_1, t_2, t_4, t_6, and t_7 represent the mail-check request, where each request lead to a burst of packet(s); the burst from t_2 to t_3 and that from t_4 and t_5 indicate packets exchanged between the client and server for email downloading. Obviously, if the proper time scale is used so that each burst (i.e., t_1, t_2 to t_3, t_4 to t_5, t_6, and t_7) is considered as the basic unit, then all these bursts together will reveal a repeating pattern (i.e., a burst appears approximately every 15 min). Otherwise, if the time scale is too small, the analysis method focuses on packets within each burst and is inclined to conclude that repeating patterns are not present.

2) Transience. Repeating patterns may appear only for a short time period. The transience challenge is commonly the result of on/off patterns of applications and hosts. Fig. 1(b) illustrates an example: an application sends packets to its server with random time intervals for most of the time and only occasionally engages in periodic packet exchange (i.e., from t_1 to t_2). Compared to the entire window of investigation, this period (from t_1 to t_2) is small. Most importantly, the boundary of repeating patterns (i.e., t_1 and t_2) is unknown a priori.

3) Interleaved by Noises. Repeating patterns could be interleaved by significant noises. This challenge commonly emerges when many applications attempt to reconnect to their servers after the current session is disrupted. Specifically, a client reconnects to its server with a gradually prolonged retry latency (i.e., sending the second reconnection packet 5 s after the first one and then sending the third packet 10 s after the second one and so on). This reconnection pattern repeats every time when the current connection is disrupted. However, once the connection is resumed, the client will exchange packets with its server whose behaviors will be dominated by random time intervals. Fig. 1(c) presents such an example: a segment of packets from t_1 to t_2 recurs from t_3 to t_4. These two segments together form a repeating pattern. It is worth noting that packets inside each pattern do not have to be repeating. However, these two segments are interleaved by a significant number of packets with random time intervals. As indicated in Fig. 1(c), packets before t_1, those from t_2 to t_3, and those after t_4 are all noises. Again, the boundary of each segment is unknown.