Future Generation Computer Systems 37 (2014) 496–511

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

PPFSCADA: Privacy preserving framework for SCADA data publishing

Adil Fahad [a,b,*], Zahir Tari [a], Abdulmohsen Almalawi [a,c], Andrzej Goscinski [d], Ibrahim Khalil [a], Abdun Mahmood [e]

[a] School of Computer Science and Information Technology, RMIT University, Melbourne, Victoria, Australia
[b] Department of Computer Science, Al-Baha University, Al-Baha City, Saudi Arabia
[c] Faculty of Computing and IT King Abdulaziz University, Jeddah, Saudi Arabia
[d] Deakin University, School of Information Technology, Melbourne, Australia
[e] New South Wales University, School of Engineering and IT, Canberra, Australia

HIGHLIGHTS

- Propose privacy-preserving framework for SCADA data publishing.
- Propose three similarity measurements to deal with multivariate network attributes.
- Propose a SCADA platform to provide real-time communication with external devices.
- Compare the proposed PPFSCADA against existing privacy-preserving approaches.

ARTICLE INFO

Article history:
Received 26 December 2012
Received in revised form
3 July 2013
Accepted 8 March 2014
Available online 20 March 2014

Keywords:
Privacy preservation
Data publishing
Data security
SCADA
Internet

ABSTRACT

Supervisory Control and Data Acquisition (SCADA) systems control and monitor industrial and critical infrastructure functions, such as electricity, gas, water, waste, railway, and traffic. Recent attacks on SCADA systems highlight the need for stronger SCADA security. Thus, sharing SCADA traffic data has become a vital requirement in SCADA systems to analyze security risks and develop appropriate security solutions. However, inappropriate sharing and usage of SCADA data could threaten the privacy of companies and prevent sharing of data. In this paper, we present a privacy preserving strategy-based permutation technique called PPFSCADA framework, in which data privacy, statistical properties and data mining utilities can be controlled at the same time. In particular, our proposed approach involves: (i) vertically partitioning the original data set to improve the performance of perturbation; (ii) developing a framework to deal with various types of network traffic data including numerical, categorical and hierarchical attributes; (iii) grouping the portioned sets into a number of clusters based on the proposed framework; and (iv) the perturbation process is accomplished by the alteration of the original attribute value by a new value (clusters centroid). The effectiveness of the proposed PPFSCADA framework is shown through several experiments on simulated SCADA, intrusion detection and network traffic data sets. Through experimental analysis, we show that PPFSCADA effectively deals with multivariate traffic attributes, producing compatible results as the original data, and also substantially improving the performance of the five supervised approaches and provides high level of privacy protection.

1. Introduction

SCADA (Supervisory Control and Data Acquisition) systems control and monitor industrial and critical infrastructure functions, such as electricity, gas, water, waste, railway, and traffic [1–3]. In the past, such systems for controlling the national critical infrastructures were inherently secure as they had proprietary controls and limited connectivity. However, the increased connectivity to the Internet and corporate networks has infinitely expanded the ability of outsiders to breach security. Nevertheless, due to the

* Corresponding author at: School of Computer Science and Information Technology, RMIT University, Melbourne, 3001 Victoria, Australia. Tel.: +61 3 9925 3782.
E-mail addresses: aalharthi.ahmed@gmail.com, s3125087@rmit.edu.au (A. Fahad), zahir.tari@rmit.edu.au (Z. Tari), abdulmohseen.almalwi@rmit.edu.au (A. Almalawi), andrzej.goscinski@deakin.edu.au (A. Goscinski), Ibrahim.khalil@rmit.edu.au (I. Khalil), Abdun.Mahmood@unsw.edu.au (A. Mahmood).

http://dx.doi.org/10.1016/j.future.2014.03.002
0167-739X/

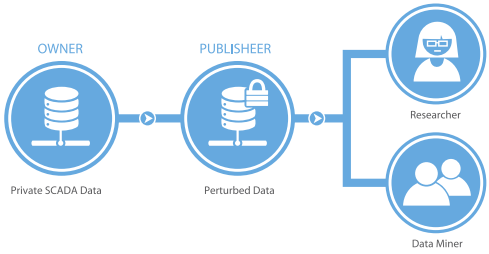

Fig. 1. Data collection and data publishing.

use of commodity hardware, software, and standard protocols [4], SCADA systems are no longer immune to cyberattacks. In fact, the threat posed to critical infrastructure is far greater in terms of impact and scale of attack than common computer vulnerabilities. Examples of threats to SCADA include an attack on a SCADA-run sewage plant in Maroochy Shire, Queensland, causing 800,000 l of raw sewage to be released into local parks and rivers, resulting in the death of local marine life as well as discoloring the water and generating a noxious stench that permeated the air [5], and the Davis–Besse nuclear power plant in Oak Harbor, Ohio, was attacked by the Slammer SQL server worm, which disabled a safety monitoring system of the nuclear power plant for nearly 5 h [6]. More recently, Stuxnet [7], a threat specifically written to target industrial control systems was discovered. The threat was designed to damage nuclear power plants in Iran [8,9]. Such utilities are essential in the proper functioning of our daily life because of the material consequences they can have to people and the environment. Therefore, it is important to analyse the security risks and develop appropriate solutions to protect such systems from malicious attacks [2].

Recently, intrusion detection systems (IDs) based on machine learning techniques [10–12] have been proposed to assist network experts to analyse the security risks and detect attacks against their systems. An extensive survey of a SCADA-based IDSs using machine learning can be found in [13]. The necessity of using machine-learning techniques for intrusion detection systems is due to their ability, to cope with a vast amount of historical data, as it is difficult for human beings to infer underlying traffic patterns in such an enormous amount of data. However, one of the key problems in the design of any SCADA-based IDS (using ML) is the lack of SCADA data [14,15]. Unfortunately, such data is not so easy to obtain, because organizations do not want to reveal their private traffic data for various privacy, security and legal reasons [15–17]. Most organizations often (and even systematically) do not want to admit that they were attacked and therefore are unwilling to give any information on this. It is therefore widely recognized today that SCADA data confidentiality and privacy are increasingly becoming an important aspect of data sharing and integration [17,18].

1.1. Contribution

This paper proposes a new privacy-preserving data framework to facilitate SCADA data publishing while ensuring that private data will not be disclosed. Fig. 1 describes a typical scenario for the data collection phase and publishing phase. In the former phase, a data publisher collects the data from the record owner (SCADA companies/organizations). In the latter phase, the data publisher releases the transformed data to a data miner or to the public, called a data recipient, who will then conduct data mining on the published data. The contributions of this paper can be summarized as follows:

- a privacy-preserving framework (PPFSCADA) based on a permutation technique is proposed to deal with SCADA traffic data. Though the vast majority of existing approaches (e.g [19–21]) on privacy-preserving computation have been active in other domains, including marketing data and biomedical data, such studied schemes are not readily applicable to private data in SCADA networks. This is mainly because they assume that the data being protected have to be numerical. A key challenge with SCADA traffic data is the need to deal with various types of attributes, such as numerical attributes (with real values), categorical attributes (with unranked nominal values), and attributes with a hierarchical structure. For example, byte counts are numerical, protocols are categorical, and IP addresses have a hierarchical structure [22]. Consequently, the proposed PPFSCADA framework is proposed to satisfy the privacy requirements while maintaining sufficient data utility. First, the traffic mixed data set is subdivided into the attributes of flow record, creating fragments: *pure categorical data set*, *pure numerical data set* and *pure hierarchical data set*. Next, well-established similarity measures to deal with various types of attributes are used to help produce more meaningful clusters. Last, the clustering results on the numerical, categorical as well as hierarchical data sets are combined as a categorical data set, on which the ML (machine learning) classifiers are employed to obtain the final output. In particular, the objective of such a framework is to enforce privacy-preserving paradigms, while minimizing the information loss incurred in the anatomizing process;
- industrial control system security (SCADA) has been a topic of scrutiny and research for several years, and many security issues are well known. However, a key challenge in the research and development of security solutions for SCADA systems is the lack of proper modeling tools due to the fact that it is impractical to conduct security experiments on a real system because of the scale and cost of implementing standalone systems. The second contribution of this paper is the development of a SCADA platform to provide a modular SCADA modeling tool that allows real-time communication with external devices using SCADA protocols. Such a platform is important not only to evaluate (i) our proposed privacy-preserving framework (PPFSCADA), but also (ii) enabling an additional benefit of testing real attacks and trying different security solutions for such systems;
- the proposed PPFSCADA framework is evaluated on both synthetic and real-life data sets. In particular, we compare the effectiveness of the PPFSCADA against a new class of privacy-preserving data mining approaches, namely: *PCA–DR* [20], *SDP* [19] and *RDP* [19]. A general observation indicates that the proposed framework outperforms the existing approaches with respect to a comprehensive set of criteria including: *Dealing With Multivariate Data*, *Efficiency*, *Scalability*, *Data Quality* and *Privacy level* (see Section 5 for details).

1.2. Organization of the paper

The rest of the paper is organized as follows: Section 2 provides a literature review of this research field. Section 3 introduces PPFSCADA as a new technique for privacy-preserving data publishing. In Section 4, we describe our SCADA platform, its main components, and data processing. In Section 5, we evaluate the performance of PPFSCADA in anatomizing the SCADA and network traffic data sets. In Section 6, we conclude the paper and discuss future research.

2. Related work

Intrusion Detection Systems (IDSs) for SCADA systems have drawn significant attention over the past few years [23,13,24,25]. IDSs usually fall into two separate categories: (i) misuse detection approach (which detects only known attacks) [23,13], and