Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 114 (2017) 149–158

Procedia Computer Science

www.elsevier.com/locate/procedia

Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems, CAS October 30 – November 1, 2017, Chicago, Illinois, USA

Data Swapping for Private Information Sharing of Web Search Logs

Kato Mivule

kmivule@nsu.edu

Department of Computer Science, Norfolk State University, Norfolk, Virginia, USA

Abstract

With the increasing number of sophisticated cyber attacks on both government and private infrastructure, cybersecurity data sharing is critical for the advancement of collaborative research among various entities, both in government, private sector, and academia. Of recent, the US Congress passed the Cyber Intelligence Sharing and Protection Act, as a framework for data sharing between various entities. Nevertheless this development raises the issue of trust between the collaborating parties, since shared data could be revealing. Conversely, due to the sensitive and confidential nature of the data involved, entities would have to employ various anonymization techniques to meet legal requirements in compliance with confidentiality policies of both their own organizations and federal government requirements. Secondly, a basic sharing of the data without the privatization process could make entities involved vulnerable to insider and inference attacks. For instance, an entity sharing data on cyber attacks might accidently reveal a sensitive network topology to an untrusted collaborator. As a contribution, we propose a modest but effective data privacy enhancement heuristic; a targeted *2k* basic data swapping of individual web search log records. In this heuristic, if individual has a set of *x* records in their web search log set *A*, those records are swapped in that individual set *A*, then swapped again with another individual *y* records in set *B*. Our preliminary results show that data swapping is effective for big data and it would be demanding to trace the original issuer of the queries in a given large dataset of web search logs, thus providing some level of confidentiality.

Peer-review under responsibility of the scientific committee of the Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems.

Keywords: Cybersecurity, Information Sharing; Data Swapping; Data Privacy; Information Security

1. Introduction

With the rising number of complex cyber attacks on both government and private infrastructure, cybersecurity data sharing is vital for the progression of collaborative research among diverse entities, both in government, private sector, and academia. To tackle this problem, the United States Congress passed the H.R.234 - Cyber Intelligence Sharing and Protection Act, which stipulates a legal framework for both government and the private sector to collaborate and

1877-0509
Peer-review under responsibility of the scientific committee of the Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems.
10.1016/j.procs.2017.09.017

share data on cybersecurity [1]. However this development raises the issue of trust between the collaborating parties, since shared data could be revealing. Conversely, due to the sensitive and confidential nature of the data involved, entities would have to employ various anonymization techniques to meet legal requirements in compliance with confidentiality policies of both their own organizations and federal government requirements. Secondly, a basic sharing of the data without the privatization process could make entities involved vulnerable to insider and inference attacks. For instance, an entity sharing data on cyber attacks might accidently reveal a sensitive network topology to an untrusted collaborator. In this study, special emphasis is placed on web search log data. The motivation for specifically targeting web search log data, is that web searches have become ubiquitous and part of the daily online search experience of individuals and organizations. This means that millions of web log records generated from Internet accessible devices, often with sensitive information detailing the intent of individuals are gathered and stored by search engines and Internet service providers (ISP). Such web log data could include indicators that would be of interest for instance, to law enforcement and advertisement entities. Yet such data cannot be freely shared among research collaborators, as this would be too revealing. Furthermore, online searches made by organizations could also be revealing in regards to intellectual property and research made. For instance, a search engine or ISP could make a good prediction about what patent or research paper a particular organization intends to write about based on the web search logs made by that organization. Therefore a naive sharing of such web log data would make individuals and organizations vulnerable to privacy breaches and potential loss of intellectual property. Such a scenario is not far fetched but was clearly highlighted by the 2006 AOL data scandal in which researchers and journalists were able to reconstruct the full identify of an individual from the naively published AOL web search logs [2][3]. Yet still, the scenario of not sharing any web search log data would only impend any research efforts that may generate novel measures to counter cybersecurity threats. It is in this light that we utilize data swapping, a data privacy technique first proposed by Dalenius and Reiss [4]. Data swapping has been favorably used by the US Census Bureau and the UK government, although the details of such algorithms have been kept secret, and as such, not much research has been done on the implementation and redesigning of the data swapping algorithms [5]. While data swapping has been shown to be a strong privacy mechanism, usability of the privatized data remains a challenge [6]. We therefore present a data swapping heuristic as a privacy preserving mechanism while seeking to maintain the usability of the privatized data, in this case, web search logs. In this heuristic, if individual has a set of x records in their web search log set A, those records are swapped in that individual set A, then swapped again with another individual y records in set B. Our preliminary results show that this modest but effective heuristic would be difficult to trace the original issuer of the queries in a given large set of web search logs. Our assumption is that users generate millions of web search logs each given day, given the ubiquitous use of the Internet as a source of information. These web search logs could be generated from smartphones, web browsers on local PCs, Tablets, and other Internet of things (IoT) devices. For instance, Google alone records a whopping 3.5 billion web searches every day [7]. Assuming that all these web search logs are stored, it would be feasible and practical to implement the proposed web search log privacy heuristic. The sheer volume of data would provide an additional cover on top of the suggested privacy implementation.

2. Background

2.1. Data swapping

Data swapping was proposed by Dalenius and Reiss (1978), as a data privacy mechanism that involves the exchange values in a variable from the same dataset while preserving the original statistical traits of the data in that variable [4][8]. Data swapping techniques maintains original statistical traits of data and therefore are used favourably by the US Census Bureau [9]. However, the US Census Bureau and other organizations keep their data swapping algorithms secret and as such, not much study and literature exists on data swapping as a data privacy mechanism [5]. Furthermore, researchers have observed the problem of data usability in data swapping since swapping deforms data by altering the joint distributions between swapped and non-swapped attributes [10]. In this study, the proposed data swapping heuristic bases on the work of Dalenius and Reiss (1978), to implement data swapping for web search logs. Furthermore, the assessment of this study is that since data swapping algorithms are guarded as a treasure by organizations, then an investigative approach to data swapping based on the work of Dalenius and Reiss becomes reasonable and foundational in generating new privacy preserving algorithms.