Journal of Computer Science 3 (11): 874-881, 2007
ISSN 1549-3636
© 2007 Science Publications

Customized Privacy Preservation Using Unknowns to Stymie Unearthing Of Association Rules

J.Indumathi and G.V. Uma
Department of Computer Science and Engineering, Anna University,
Chennai-600 025.Tamilnadu, India

Abstract: The explosions of new data mining techniques has augmented privacy risks because now it is probable to powerfully coalesce and cross-examine massive data stores, accessible on the web, in the rummage around of earlier unidentified hidden patterns. Consecutively to make a overtly accessible system safe and sound, we must guarantee not only that private sensitive data have been trimmed out, but also to make certain that certain inference channels have been clogged-up. The data and the concealed knowledge in this data should be made secure. Furthermore, the requirement for making our system as open as probable - to the extent that data sensitivity is not jeopardized - asks for diverse techniques that account for the revelation organize of sensitive data. At its nucleus, the value of privacy preserving data mining is plagiaristic not only from its knack to haul out imperative knowledge, but also from its resiliency to molestation. It performs well at needed levels during times of both crisis and normal operations. This task force's central thrust is towards establishing a earth with robust data security, where knowledge users persist to profit from data without compromising the data privacy.The goal of privacy-preserving data mining is to liberate a dataset that researchers can study without being able to identify sensitive information about any individuals in the data (with high probability). One technique for privacy-preserving data mining is to replace the sensitive items by unknown values. For many situations it is safer if the sanitization process consign unknown values as a substitute of fake values. This obscures the susceptible rules, whilst defending the punter of the data commencing false rules. In this study, we modify the blocking algorithms of[1] by proposing a new heuristic in order to reduce the information loss. We put forward an enhanced approach that overcomes the privacy breach problem of existing blocking approaches. Though they have argued that the rules are truly safe from an attack by an adversary, they have not formally proved the safety, which we have proved. We have investigated how probabilistic and information theoretic techniques can be applied to this problem. More complete analysis of the effectiveness of these rule obscuring techniques, and formal study of the problem has been made. Our preliminary domino effect point toward deterministic algorithms for privacy preserving association rules shows potential framework for controlling disclosure of sensitive data and knowledge.

Key words: Adversary, confidence, data sanitization, disclosure control, inference problem, machine learning, network, repository, support

INTRODUCTION

Insecurity of Computers has become a worldwide observable fact. scattered refutation of service attacks, speedily propagating viruses, self-replicating worms are a nuisance of computer networks global, and attacks continuously breed in brutality and erudition. Subsequent to the trendy success of initiatives as DShield[4] and DeepSight[16], there has been a mounting interest in the construction of large-scale analysis centers that bring together network security information starting a sundry pool of contributors and afford a real-time warning service for Internet threats.

Availability of rich, ample datasets composed from a extensive cross-section of intrusion detection systems, fire- walls, honey pots, and network sensors has the prospective to reason a exemplary shift in computer security research.

It has been renowned; nonetheless, that open access to raw system security data is burdened with threat. Even lawful access to the data can be battered,

Corresponding Author: J. Indumathi, Department of Computer Science and Engineering, Anna University, Chennai- 600025. Tamilnadu, India

and the data contributed by well-intentioned mutual partners can be turned against them. For example, security alerts contributed by network sensors can be used to finger- print these sensors and to map out their locations[3].

Over the last few decades, there has been a budding concern in the development of wide-area data collection and investigation centers to help identify, track, and formulate responses to the ever-growing number of coordinated attacks and malware infections that plague computer networks worldwide. As all-embracing network threats protract to expand in erudition and extend to expansively deployed applications, we foresee that concern in mutual security monitoring infrastructures will continue to nurture, because such attacks may not be effortlessly diagnosed from a single point in the network. We tried to outline the prominent issues faced by network security centers, review proposed defense mechanisms, and pretense several research challenges to the computer security community.

Data mining is the process of extracting valuable and lucid information from raw data. This field is cross-disciplinary in temperament and draws from research in statistics, machine learning, and databases. Data mining techniques helps business people to take intelligent decisions by mining interesting knowledge from huge databases. But there is some sensitive information that is not to be mined; hence a proper balance of privacy and mining has become essential.

LITERATURE SURVEY

At its nucleus, the value of privacy preserving data mining is plagiaristic not only from its knack to haul out imperative knowledge, but also from its resiliency to molestation. It performs well at needed levels during times of both crisis and normal operations. This task force's central thrust is towards establishing a earth with robust data security, where knowledge users persist to profit from data without compromising the data privacy.The goal of privacy-preserving data mining is to liberate a dataset that researchers can study without being able to identify sensitive information about any individuals in the data (with high probability). One technique for privacy-preserving data mining is to replace the sensitive items by unknown values. For many situations it is safer if the sanitization process consign unknown values as a substitute of fake values. This obscures the susceptible rules, whilst defending the punter of the data commencing knowledge \false rules.

The technique offered here applies to applications where it is obligatory to stock up woolly or unknown values for some attributes, such as when authentic values are kept a secret or unavailable. We offer the technique for hiding rules (i.e., knowledge) from a data set, by replacing select attribute values with unknowns. This is similar to previous proposals that replace select values with \false values[13]. The fake values is capable of boast ghastly consequences. Consider a medical body that will make some of its data open, and the data is sanitized by replacing real attribute values by fake values. Researchers may use this data, but attain disingenuous results In the worst case, such misleading data could be used for critical purposes (like analysis) and jeopardize patients' lives. at the same time as a outcome, for many situations it is safer if the sanitization process consign unknown values as a substitute of fake values.

In order to safeguard privacy while disclosing data sets, it must be sure fire that merely secret data values and private knowledge within the data set which is discoverable by certain data mining methods should be concealed. Correlations or set of laws are examples of such secret information that one can gain knowledge from data. The predicament of suppressing secret association rules has been addressed in literature[14,18,10]. These approaches utilize data sanitization both by distorting or blocking the data sets. The distortion approaches[2,3] introduce false values to the data sets to avoid discovery of confidential association rules. However, these false values reduce the trustworthiness and usefulness of the data sets.

Blocking approaches[1] defeat this problem by introducing unknown values denoted by "?" to the data sets. besides doing so, they restrain a specified set of confidential rules whereas plummeting the side effects on non-confidential ones. Nevertheless, these approaches too contain certain drawbacks. First of all, they do not try to diminish the information loss. Furthermore, they cause privacy breaches on top of the customized data set. In reality, knowing the blocking approach engaged, an antagonist can infer authentic values of all? s and ascertain the confidential association rules. In this study, we modify the blocking algorithms of[1] by proposing a new heuristic in order to reduce the information loss. We put forward an enhanced approach that overcomes the privacy breach problem of existing blocking approaches. Though they have argued that the rules are truly safe from an attack by an adversary, they have not formally proved the safety, which we have tried to achieve. We have investigated how probabilistic and information theoretic techniques can also be applied to this problem. More complete analysis of the effectiveness of these rule obscuring techniques, and formal study of the problem has been made.

Fortification and sanitization of network security data has established some interest in the precedent years[6,7,15,19]. The objective of this study is to formulate several brusque research challenges for the computer security society. We deem that these challenges will inspire the debate, spur design and implementation of efficient sanitization technologies that balance the utility of network security data for mutual analysis against the need to protect contributors' privacy and security, and even escort to new paradigms for large-scale sharing of network data, including security alerts, packet traces, and so on.

Risks and challenges can be classified into three areas of concern, viz., network sensors that generate the data, repositories that collect the data and make them available for analysis, and the network infrastructure which delivers the data from sensors to repositories.

The classes of threats referred to as fingerprinting attacks on network data have proved overwhelmingly effective in many contexts[3,5]. In a fingerprinting molestation, an attacker may search for normal patterns in the data that distinctively identify a fastidious host (e.g., clock skew[5]). On the other hand, the assailant may aggressively sway data patterns by triggering rare rules in signature-based intrusion detection systems, employing atypical port combinations, or generating definite event sequences or timing patterns that can afterward be improved from the depository. (i.e., known in the text as the probe response attack[6,3].) Probe response and fingerprinting attacks twist the accustomed intrusion detection game on its head. Here, the attacker's goal is to dodge detection, and he wants to be detected so that he canister scrutinizes the resultant testimony for substantiation of vulnerabilities and gain improved indulgent of the defender's security posture. painstaking formalization of fingerprinting attacks and improvement of provably safe and sound defense mechanisms alongside fingerprinting are in the midst of the most vital challenges branded in this study. Privacy-preserving alteration and anonymization of Internet packet traces[11,12,17] and routing configuration data[8] have established a lot of interest in the network research community.

PROBLEM DESCRIPTION FOR CUSTOMIZED PRIVACY PRESERVATION USING UNKNOWNS

In this research, we modify the blocking algorithms of[1] by proposing a new heuristic in order to reduce the information loss. We put forward an enhanced approach that overcomes the privacy breach problem of existing blocking approaches. Though they have argued that the rules are truly safe from an attack by an adversary, they have not formally proved the safety, which we have proved. We have investigated how probabilistic and information theoretic techniques can be applied to this problem. More complete analysis of the effectiveness of these rule obscuring techniques, and formal study of the problem has been made. Our preliminary domino effect point toward deterministic algorithms for privacy preserving association rules shows potential framework for controlling disclosure of sensitive data and knowledge. The idea is to have a repository somewhat loosely to denote both open and restricted-access analysis centers, which collect network security data from contributors and make it available either in raw, or in sanitized form. Repository of data becomes a single point of failure and a natural target for attackers, not to mention insider compromise. Moreover, even legitimate access to the data can be abused, and the data contributed by well-intentioned collaborative partners can be turned against them. For example, security alerts contributed by network sensors can be used to finger- print these sensors and to map out their locations[5]. Security and audit logs may passively leak information about the contributor's vulnerabilities, as well as the data about topology of protected networks, enabled services and applications, egress filtering policies, and so on.

PROPOSED ARCHITECTURE

One must redact information as of the raw data prior to providing it. The redaction may possibly get one of two forms: *summary* or *sanitization*.

A summary is a finished analysis of the data in which the pertinent information is used to calculate statistics for instance counts, means, and so forth. Characteristically, a gross depiction of the data is well-known to provide milieu. The Indian Bureau of Statistics presents summaries of raw data to the public, and does so in such a way that the raw data cannot be redefined from the synopsis. Consequently the summary conceals all aspects of the raw data that the government department desires to restrain. A summary of the communal data mentioned above would name the state the quantity of houses purchased by clients from that state.

Sanitization: Takes the contradictory approach. The raw data is offered for others to analyze, however the data is altered so that sensitive items are suppressed.

Medical records given to researchers are treated in this style. Diagnostic information, symptoms, and treatments are given, but information identifying the exact patient, such as name, address, phone number, and so forth, are redacted. A sanitized set of the corporate data mentioned above might consist of a list of purchasers, their addresses, and the number of houses each purchased, but the names and addresses would be replaced by meaningless strings.The remuneration of sanitization are that the beneficiary of the data can analyze the raw data and obtain statistics, or take activities, based upon the data itself sooner than the provider's summary of the data. Based upon the components of the data that are sanitized the recipient needs to derive information which is a serious setback.

Based on the total set of data to be sanitized accessible time

Static sanitization: When the total set of data to be sanitized is accessible at the time of sanitization, the sanitization functions can be derived completely prior to the sanitisation of the data.

Dynamic sanitization: When the total set of data to be sanitized is not obtainable at the time of sanitization, the sanitization function may change as the data becomes available and is sanitized.

Basic ways to sanitize objects are[9]:

- **Deletion.** Here, the objects to be sanitized are minimally deleted.
- **Fixed transformation.** All occurrences of the object are replaced by a fixed string.
- **Variable transformation.** Occurrences of the object are transformed in different ways depending upon the context and structure of the object. For example, translating an IP address into one value for FTP connections, and a different value for HTTP messages, is an example of variable replacement. Replacing objects with random data is another.
- **Typed transformation.** This is a form of variable transformation, except that the replacing objects are related when the types of the object being replaced are the same. For example, replacing all file names with a value generated by one cryptographic hash function, and all IP addresses with addresses selected from the 10. network, would be an example of this. As shown in Fig. 4a we transform a database into a new one that conceals some premeditated patterns (restrictive association rules) while preserving the general patterns and trends from the original database. The procedure of transforming an original database into a sanitized one is called data sanitization The sanitization process acts on the data to remove or hide a group of restrictive association rules that contain sensitive knowledge.

SYSTEM ARCHITECTURE DESIGN

Data structure design and datasets used

Database Description: The accidents data are collected and stored in MsAccess.Connectivity from C# is done using ADO.Net. The support for the victim item and the other items can be displayed when the item is selected[2].

User Interface Design Specifications

GUI : The graphical user interface is designed using C#

Operating environment: Net Framework

GUI description

ComboBox1: User Threshold : User can either select 50% -partial hiding or 100% -full hiding.

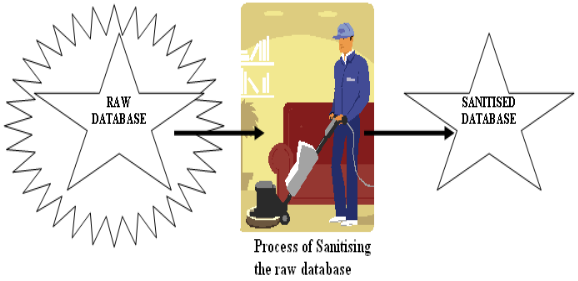


Fig. 4a: Sanitization Process of a raw database: High Level

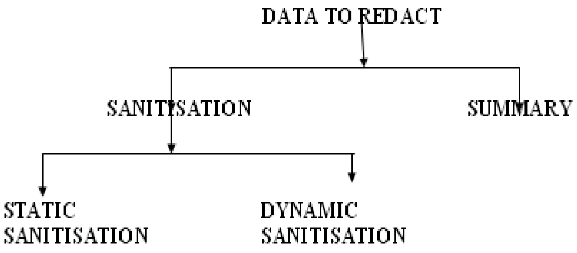


Fig. 4b: Taxonomy of data to redact

ComboBox1: Support : User can select which of the items are sanitized and their corresponding support.

Text Filed : Datawarehouse name: The name of the datawarehouse that is the sanitized.

Button1: Desanitized Database: When the user clicks this button the database before Sanitization is displayed as shown in the Fig. 5a.

Button2: Sanitized Database: When the user clicks this button the database after Sanitization is displayed as shown in the Fig. 5b.

User Form: This form has a login page which has got validation for the users**.**

IMPLEMENTATION

Customised algorithm

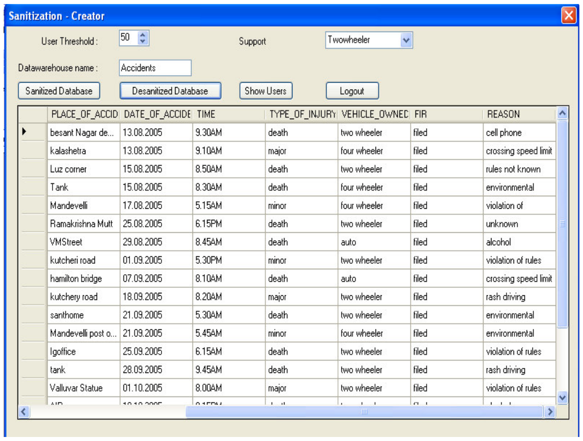
Step 1: Extract sensitive data using sensitive association rules and place the data in dataset D

Step 2: The dataset in D are arranged ascending order of their size (Where size is determined by the number of itemsets.The idea behind choosing the shortest item set for removal is that, a short transaction will possibly have less side effects on the other item sets than a long item set)

Step 3: Sort the items in each item sets in descending order or the support. (The algorithm chooses the item with highest minimum support for removal with the intention that an item of high minimum support will have less side effects since it has many more transactions that support it compared to an item of low minimum support)

Step 4: Now the selected item the victim item is replaced by?

Step 5: The support of the victim item is updated [The algorithm as shown in Fig. 6a initially retrieves all item sets that are sensitive using the rules and placed in a database D.This item sets are first sorted in ascending order because the removal of shortest item set will have less side effects compared to long itemset.In each of the item set they are arranged in descending order because removal of the item with highest minimum support will have less side effects compared to item with less support count. The item with highest support count is selected and replace by? And the support of the corresponding item is updated.



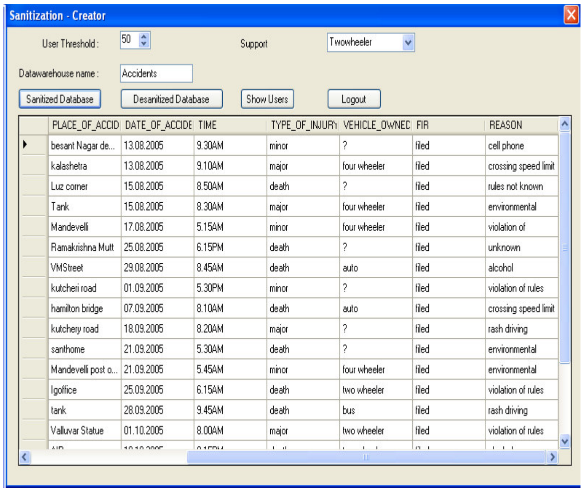
User Threshold : 50

Support Twowheeler

Datawarehouse name : Accidents

PLACE_OF_ACCID	DATE_OF_ACCIDE	TIME	TYPE_OF_INJURY	VEHICLE_OWNED	FIR	REASON
besant Nagar de...	13.08.2005	9.30AM	death	two wheeler	filed	cell phone
kalashetra	13.08.2005	9.10AM	major	four wheeler	filed	crossing speed limit
Luz corner	15.08.2005	8.50AM	death	two wheeler	filed	rules not known
Tank	15.08.2005	8.30AM	death	four wheeler	filed	environmental
Mandevelli	17.08.2005	5.15AM	minor	four wheeler	filed	violation of
Ramakrishna Mutt	25.08.2005	6.15PM	death	two wheeler	filed	unknown
VMStreet	29.08.2005	8.45AM	death	auto	filed	alcohol
kutcheri road	01.09.2005	5.30PM	minor	two wheeler	filed	violation of rules
hamilton bridge	07.09.2005	8.10AM	death	auto	filed	crossing speed limit
kutchery road	18.09.2005	8.20AM	major	two wheeler	filed	rash driving
santhome	21.09.2005	5.30AM	death	two wheeler	filed	environmental
Mandevelli post o...	21.09.2005	5.45AM	minor	four wheeler	filed	environmental
Igoffice	25.09.2005	6.15AM	death	two wheeler	filed	violation of rules
tank	28.09.2005	9.45AM	death	two wheeler	filed	rash driving
Valluvar Statue	01.10.2005	8.00AM	major	two wheeler	filed	violation of rules

Fig. 5a: Display of data before sanitization



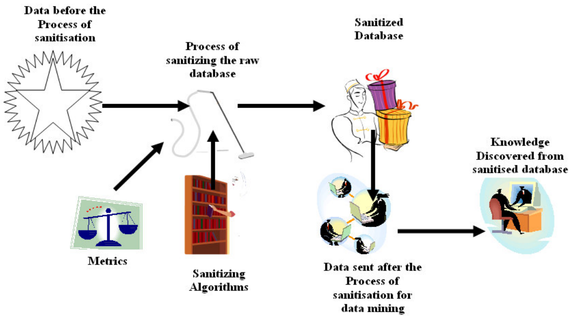
User Threshold : 50

Support Twowheeler

Datawarehouse name : Accidents

PLACE_OF_ACCID	DATE_OF_ACCIDE	TIME	TYPE_OF_INJURY	VEHICLE_OWNED	FIR	REASON
besant Nagar de...	13.08.2005	9.30AM	minor	?	filed	cell phone
kalashetra	13.08.2005	9.10AM	major	four wheeler	filed	crossing speed limit
Luz corner	15.08.2005	8.50AM	death	?	filed	rules not known
Tank	15.08.2005	8.30AM	major	four wheeler	filed	environmental
Mandevelli	17.08.2005	5.15AM	minor	four wheeler	filed	violation of
Ramakrishna Mutt	25.08.2005	6.15PM	death	?	filed	unknown
VMStreet	29.08.2005	8.45AM	death	auto	filed	alcohol
kutcheri road	01.09.2005	5.30PM	minor	?	filed	violation of rules
hamilton bridge	07.09.2005	8.10AM	death	auto	filed	crossing speed limit
kutchery road	18.09.2005	8.20AM	major	?	filed	rash driving
santhome	21.09.2005	5.30AM	death	?	filed	environmental
Mandevelli post o...	21.09.2005	5.45AM	minor	four wheeler	filed	environmental
Igoffice	25.09.2005	6.15AM	death	two wheeler	filed	violation of rules
tank	28.09.2005	9.45AM	death	bus	filed	rash driving
Valluvar Statue	01.10.2005	8.00AM	major	two wheeler	filed	violation of rules

Fig. 5.2: Display of data after sanitization



Data before the Process of sanitisation
Process of sanitizing the raw database
Sanitized Database
Knowledge Discovered from sanitised database
Metrics
Sanitizing Algorithms
Data sent after the Process of sanitisation for data mining

Fig. 6.1: Sanitization Process of a raw database:Low Level

RESULTS AND ANALYSIS

The victim data item changes dynamically. The Fig. 7a and b displays their support change before sanitization and after sanitization.

Reliability: This parameter is a non functional parameter.It is defined as "obtaining the correct output with in the given time.The reliability with respect to this application is determined to be getting the right support or confidence value with in the estimated time Fig. 7c.

Availability: This is defined as the availability of the data for the application to sanitize.For sanitizing vital is the data.If the data is not made available the application cannot perform its function Fig. 7d.

ResponseTime: This is defined as the time taken by the application to sanitize the data once the user submits the data for sanitization Fig. 7e.

Scalability: This is defined as the capability of the application to scale up from small data to larger data for the process of sanitization. The algorithm is an optimal one for medium datasets but when the interestingness measures increasing the algorithm may not be feasible Fig. 7f.

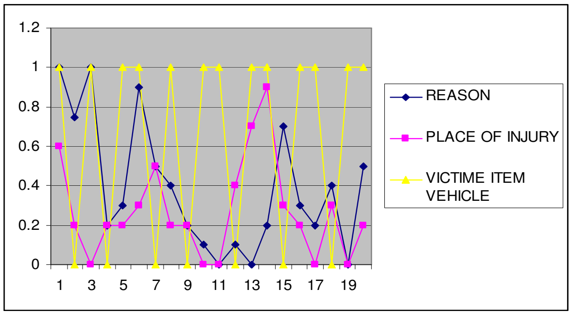


Fig. 7a: Before sanitization

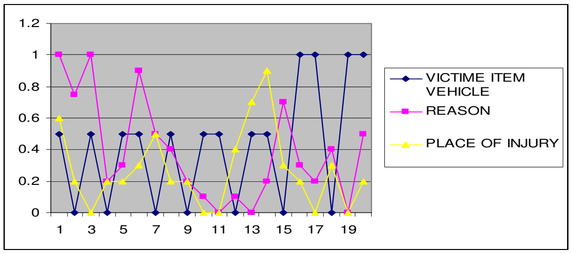


Fig. 7b: After sanitization

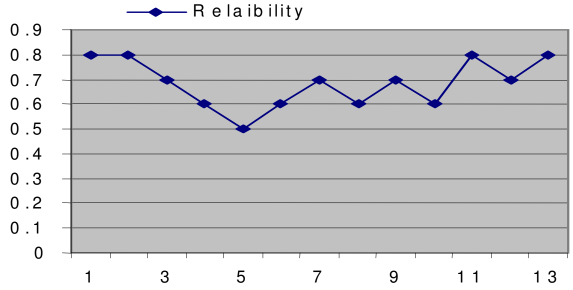


Fig. 7c: Reliability

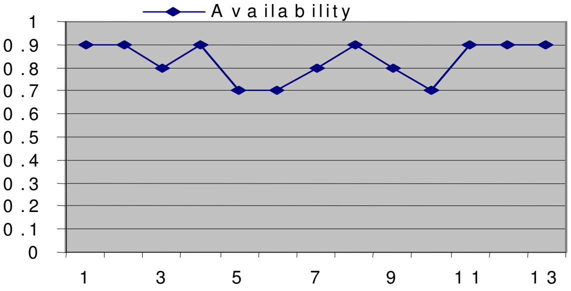


Fig.7. 4d: Availability

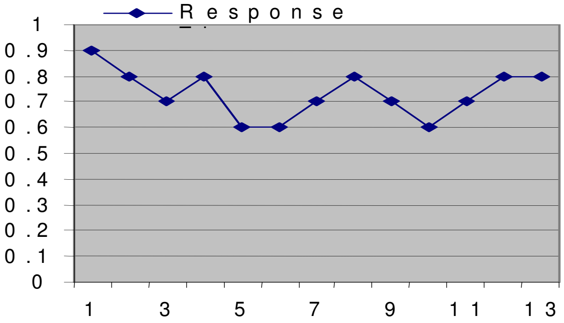


Fig. 7e: Reliability

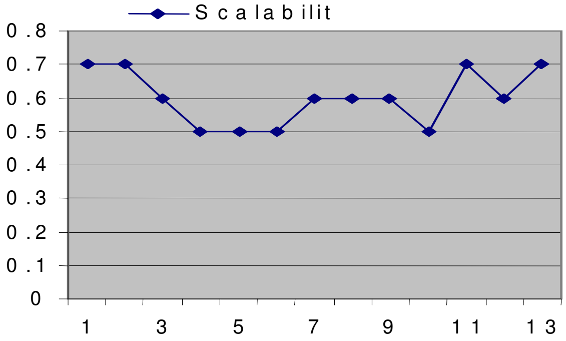


Fig. 7.6:Scalability

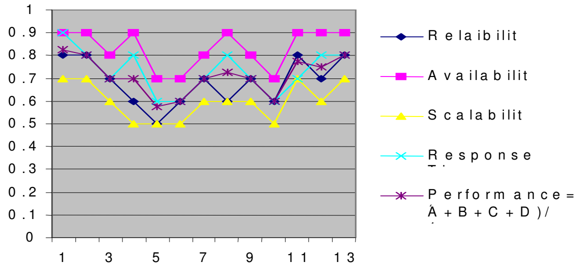


Fig. 7g: Reliability

Performance: This parameter is obtained by calculating the weighted sum of all the above non functional parameters. Reliability, Availabiltiy, Scalability needs to be high and Response Time needs to be very low. Time taken to sanitize the database in milliseconds.And this process of santization may function effectively when the item set is moderate Fig. 7g.

CONCLUSION

In recent years, as Internet attacks increased in scale, frequency, and severity, there has been a growing interest in creating global analysis centers that woul dgather net- work security data from a wide variety of network sensors, use it for real-time collaborative analysis to detect inflection points and global security trends, identify propagation patterns and attack vectors of malware, and make the data available for network security researchers. Successful deployment of global analysis centers will require resolving a number of fundamental tradeoffs between increased global network security, privacy of data contributors, potential for malicious abuse of the reported data, liability of data repositories, usefulness of the data for net- work security research, and practical efficiency.

FUTURE WORK

Rigorous formalization of finger- printing attacks to better understanding of traffic analysis attacks which de-anonymize the data contributed to global analysis centers. We hope that our challenges will become part of the research program for computer scientists working in this area. It is unlikely that global Internet defense will succeed without solving them.

REFERENCES

1. Agrawal R. and R. Srikant, 2000. Privacy-preserving data mining, In Proc. of the ACM SIGMOD Conference on Management of Data, Dallas, Texas, 439-450.
2. Back, A., U. Mo¨ller and Stiglic, 2001. A. Traffic analysis attacks and trade-offs in anonymity providing systems. In Proc. 4th International Workshop on Information Hiding , vol. 2137 of LNCS,pp. 245-257.
3. Bethencourt, J., J. Franklin and V.M. Mapping, 2005. Internet sensors with probe response attacks. In Proc. 14th USENIX Security Symposium ,pp. 193-208.
4. DShield. 2006. http://www.dshield.org, Last accessed in 2007.
5. Kohno, T., A. Broido and K. Claffy, 2005. Remote physical device fingerprinting. In Proc. IEEE Symposium on Security and Privacy, pp: 211-225.
6. Lincoln, P., P. Porras and V. Shmatikov, 2004. Privacy-preserving sharing and correlation of security alerts. In Proc. 13th USENIX Security Symposium, pp: 239-254.
7. Locasto, M., J. Parekh, A. Keromytis and S. Stolfo, 2005. Towards collaborative security and P2P intrusion detection. In Proc. IEEE Information Assurance Workshop, pp: 333-339.
8. Maltz, D., J. Zhan, G. Xie, H. Zhang, G. Hjalmtysson, A. Greenberg and J. Rexford, 2004.Structure preserving anonymization of router configuration data. In Proc. 4th ACM SIGCOMM Conference on Internet Measurement, pp: 239-244.
9. Matt Bishop, Bhume Bhumiratana, Rick Crawford, Karl Levitt. 2004. How to Sanitize Data In Proc. of the 13th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE'04)1524-4547/04.
10. Oliveira, S.R.M. and O.R. Zaiane, 2003. Protecting Sensitive Knowledge by Data Sanitization. In Proceeding of the 3rd IEEE International Conference on Data Mining, p: 613-616.
11. Pang, R. and V. Paxson, 2003. A high-level programming environment for packet trace anonymization and transformation. In Proceeding ACM SIGCOMM 03, pp: 339-351.
12. Phillip Porras and Vitaly Shmatikov. 2006. LargeScale Collection and Sanitization of Network Security Data: Risks and Challenges.In the proceedings of New Security Paradigms Workshop, Schloss Dagstuhl, Germany.
13. Verykios, S., K. Ahmed Elmagarmid, B. Elisa, Y. Saygin and D. Elena, 2004. Association Rule Hiding, IEEE Transactions on Knowledge and Data Engineering, 16(4).
14. Saygin, Y., V.S. Verykios and C. Clifton, 2001. Using Unknowns to Prevent Discovery of Association Rules. SIGMOD Record, 30(4): 45-54.

15. Slagell, A. and W. Yurcik, 2005. Sharing computer network logs for security and privacy: a motivation for new methodologies of anonymization. In Proceeding SECOVAL: The Workshop on the Value of Security through Collaboration.
16. Symantec. 2006. DeepSight threat management system.http://tms.symantec.com. Last accessed in 2007.
17. Tcpdpriv. 2006. Program for eliminating confidential information from traces. http://ita.ee.lbl.gov/html/contrib/tcpdpriv.html, Last accessed in 2007.
18. Verykios, V.S., A. Elmagarmid, E. Bertino, Y. Saygin, E. Dasseni, 2004. Association Rule Hiding. IEEE TKDE, 16(4).
19. Xu, D. and P. Ning, 2005.Privacy-preserving alert correlation: a concept hierarchy based approach. In Proc. 21st Annual Computer Security Applications Conference (ACSAC), pp: 537-546.