

Anonymization of System Logs for Privacy and Storage Benefits

Siavash Ghiasvand

Center for Information Services
and High Performance Computing
Technical University of Dresden, Germany
Email: siavash.ghiasvand@tu-dresden.de

Florina M. Ciorba

Department of Mathematics
and Computer Science
University of Basel, Switzerland
Email: florina.ciorba@unibas.ch

Abstract—System logs constitute valuable information for analysis and diagnosis of system behavior. The size of parallel computing systems and the number of their components steadily increase. The volume of generated logs by the system is in proportion to this increase. Hence, long-term collection and storage of system logs is challenging. The analysis of system logs requires advanced text processing techniques. For very large volumes of logs, the analysis is highly time-consuming and requires a high level of expertise. For many parallel computing centers, outsourcing the analysis of system logs to third parties is the only affordable option. The existence of sensitive data within system log entries obstructs, however, the transmission of system logs to third parties. Moreover, the analytical tools for processing system logs and the solutions provided by such tools are highly system specific. Achieving a more general solution is only possible through the access and analysis system of logs of multiple computing systems. The privacy concerns impede, however, the sharing of system logs across institutions as well as in the public domain. This work proposes a new method for the anonymization of the information within system logs that employs de-identification and encoding to provide sharable system logs, with the highest possible data quality and of reduced size. The results presented in this work indicate that apart from eliminating the sensitive data within system logs and converting them into shareable data, the proposed anonymization method provides 25% performance improvement in post-processing of the anonymized system logs, and more than 50% reduction in their required storage space.

I. INTRODUCTION

System logs are valuable sources of information for the analysis and diagnosis of system behavior. The size of computing systems and the number of their components, continually increase. The volume of generated system logs (hereafter, syslogs) is in proportion to this increase. The storage of the syslogs produced by large parallel computing systems in view of their analysis requires high storage capacity. Moreover, the existence of sensitive data within the syslogs raises serious concerns about their storage, analysis, dissemination, and publication. The anonymization of syslogs is a means to address the second challenge. During the process of anonymization, the sensitive information will be eliminated while the remaining data is considered as *cleansed data*. Applying anonymization methods to syslogs to cleanse the sensitive data before storage, analysis, sharing, or publication, reduces the usability of the anonymized syslogs for further analysis.

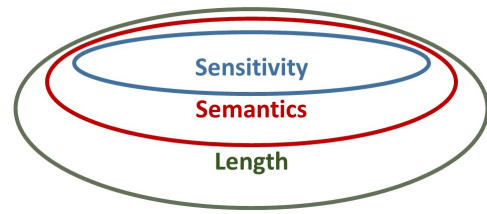


Fig. 1. The sensitivity, semantic, and length of terms in syslog entries and their relation. Each term of a syslog entry has a non-zero length. Terms may or may not have semantic. A term with semantic may or may not also be sensitive.

After a certain degree of anonymization, the cleansed syslogs lose semantic and only remain useful for statistical analysis, such as time series and distributions. At this stage, it is possible to transform long syslog entries into shorter strings. Reducing the length of cleansed and semantic-less syslogs significantly reduces the required storage capacity of syslogs and addresses the storage challenge mentioned earlier. Shortening the log entries' length reduces their processing complexity and, therefore, improves the performance of further analysis on syslogs.

In this work, we address the trade-off between the sensitivity and the usefulness of the information in anonymized syslogs. It is important to note that the sensitivity and the semantic of syslogs are relative terms. Each data item (or term) in a syslog entry, depending on policies of the computing system it originates from, may or may not be considered sensitive data. The same degree of relativity applies to the semantic of a syslog entry data item. Depending on the chosen data analysis method, the semantic of syslogs can be assessed as rich or poor. Even though the classification of each term as sensitive or as semantic is related to the policies of computing centers, the final assessment of sensitivity and semantic has a binary value of *true* (1) or *false* (0). Therefore, every single term in a syslog entry can only be sensitive or nonsensitive, e.g., a username. Fig. 1 illustrates the relation between the sensitivity, the semantic and the length of the syslog terms.

A triple trade-off exists between sensitivity, semantic, and length of a syslog entry. Fig. 2 schematically illustrates this trade-off, regardless of the system policies and syslog analysis

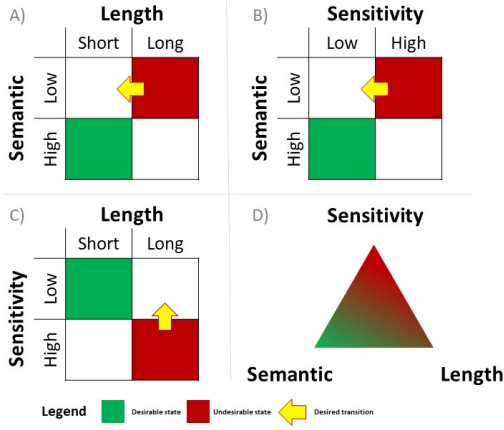


Fig. 2. Trade-off scenarios between the semantic, sensitivity, and length of a system log entry. Each of the A), B), and C) illustrations depicts the four possible states of a syslog entry based on its sensitivity, semantic, and length. These states are not limited to syslog entries alone and can be generalized to higher granularities of syslog data. The green state denotes the best situation, while the red state is most undesirable. In red states, the intention is transition to the state shown by the yellow arrow in the trade-off diagrams. The trade-off triangle in illustration D) shows the trade-off between the three parameters (sensitivity, semantic, length) in a single unified view.

methods in use. The illustration shows that a syslog entry can be in four distinct states. Green color states denote best conditions while red color states denote undesirable conditions. White color states represent neutral conditions. Under undesirable conditions, the approach taken in this work is to transition from the red state to one of the white states. The yellow arrows indicate this in Fig. 2. Increasing the semantic of a syslog entry is not possible. Therefore, the remaining possibilities are either decreasing the sensitivity or reducing the length of the syslog entry.

Data, in general, has a high quality when it is fit for [its] intended uses in operations, decision making, and planning [1]. The syslog entries represent the data in this work and several parameters affect their quality. To measure and maximize this quality, a utility function called *quality* (Q_E) is defined as the relation between sensitivity, semantic, length, and usefulness of syslog entry E . The goal of this work is to maintain the *quality* (Q_E) of all syslog entries, by pushing the parameters mentioned above toward their best possible values, when the computing system policies degrade this quality. The main contribution of this work is in introducing a new approach for anonymization that employs de-identification and encoding to provide shareable system logs, with the highest possible data quality and of reduced size.

The remainder of this work is organized as follows. In Section II the background and current state of the art are discussed. The proposed approach is described in Section III, and the methodology and technical details are provided in Section IV. After explaining the results of the current work in Section V, the conclusion and future work directions are discussed in Section VI.

II. RELATED WORK

In July 2000, the European Commission adopted a decision recognizing the "Safe Harbor Privacy Principles" [2]. Based on the "Safe Harbor" agreement, eighteen personal identifiers should be eliminated from the data before its transmission and sharing. "Safe Harbor" was originally designed to address the privacy of healthcare-related information. However, its principles are also taken into account for other types of information.

Later, in March 2014, European Parliament approved the new privacy legislation. According to this regulations, *personal data* is defined as "any information relating to an identified or identifiable natural person ('data subject.')" [3]. This information must remain private to ensure a person's privacy. Based on this definition, syslog entries contain numerous terms which represent personal data and must, therefore, be protected.

Protection of personal data in syslog entries can be attained via various approaches; the most common ones are encryption and de-identification. Encryption reduces the risk of unauthorized access to personal data. However, the encrypted syslog entries cannot be freely used or shared in the public domain. The risk of disclosure of the encryption-key also remains an important concern. In contrast, de-identification eliminates the sensitive data and only preserves the nonsensitive (cleansed) data. As such, de-identification provides the possibility of sharing de-identified information in the public domain. The de-identified data may turn out to no longer be of real use.

Pseudonymization and anonymization are two different forms of de-identification. In pseudonymization, the sensitive terms are replaced by dummy values to minimize the risk of disclosure of the *data subject* identity. Nevertheless, with pseudonymization the *data subject* can potentially be re-identified by some additional information [4]. Anonymization, in contrast, refers to protecting the user privacy via irreversible de-identification of personal data.

Several tools have been developed to address the privacy concerns of using syslog information. Most of these tools provide log encryption as the main feature, while certain such tools also provide de-identification as an additional feature. Syslog-ng and Rsyslog are two open-source centralized logging infrastructures that provide *out of the box* encryption and message secrecy for syslogs, as well as de-identification of syslog entries [5], [6]. Both tools provide a *pattern database* feature, which can identify and rewrite personal data based on pre-defined text patterns. Logstash [7] is another open-source and reliable tool to parse, unify, and interpret syslog entries. Logstash provides a text filtering engine which can search for the text patterns in live streams of syslog entries and replace them with predefined strings [8]. In addition to the off-line tools, such as Syslog-ng and Logstash, there is a growing number of on-line tools, e.g., Loggy [9], Logsign [10], and Scalyr [11], that offer a comprehensive package of syslog analysis services. The existence of sensitive data in the syslogs, barricades the usage of such services.

Alongside these industrial-oriented tools, several research

groups have developed scientific-oriented toolkits to address the syslog anonymization challenge. eCPC toolkit [12], sdcMicro [13], TIAMAT [14], ANON [15], UTD Anonymization Toolbox [16], and Cornell Anonymization Toolkit [17] are selected examples of such toolkits. These tools apply various forms of *k-anonymity* [4] and *l-diversity* [18] to ensure data anonymization. Achieving an optimal *k-anonymity* is an NP-hard problem [19]. Heuristic methods, such as k-Optimize, can provide effective results [20].

The main challenges of using existing anonymization approaches, in general, are: (1) The quality of the anonymized data dramatically degrades, and (2) The size of the anonymized syslogs remains almost unchanged. The industrial-oriented approaches are unable to attain full anonymization at micro-data [4] level. Even though scientific-oriented approaches can guarantee a high level of anonymization, they are mainly not capable of applying effective anonymization in an online manner. Certain scientific-oriented methods, such as [21], which can effectively anonymize online streams of syslogs, need to manipulate log entries at their origin [22].

The anonymization approach proposed in this work is distinguished from existing work through the following features: (1) Ability to work with streams of syslogs without modification of the syslog origin; (2) Preservation of the highest possible quality of log entries; and (3) Reduction of the syslogs storage requirements, whenever possible.

III. PROPOSED APPROACH

Computing systems can generate system logs in various formats. RFC5424 proposes a standard for the syslog protocol which is widely accepted and used on computing systems [23]. According to this protocol, all syslog entries consist of two main parts: a *timestamp* and a *message*. In addition to these two main parts, there are optional parts, such as *system tags*. Let us consider the following sample syslog entry E_1 : "1462053899 Accepted publickey for Siavash from 4.3.2.1". In this entry, "1462053899" is the *timestamp* and the rest of the line "Accepted publickey for Siavash from 4.3.2.1" is the *message*. In the *message* part, the terms Accepted, publickey, for, and from are *constant* terms, while Siavash, and 4.3.2.1 are *variable* terms, in the sense that for the above constant terms, the user name and IP can vary among users and machines.

The goal of this work, described earlier in Section I, is to preserve the *quality* of syslog entries throughout the anonymization process. To achieve this goal, (1) The *variable* terms in the syslog entries are divided into 3 groups: *sensitive*, *meaningful* (those that have a semantic), and *semantic-less* terms. (2) The *sensitive* terms are eliminated to comply with the privacy policies. (3) The *semantic-less* terms are replaced with predefined constants to reduce the required storage. (4) Following the anonymization steps (2) and (3) above, every syslog entry that does not have any additional *variable* terms, is mapped to a hash-key, via a collision-resistant hash function. This step is called *encoding*. (5) The quality of the remaining syslog entries is measured with a utility function.

(6) When it is revealed that removing a *meaningful* term from the syslog entry improves the quality of syslog, that particular term is replaced with a predefined constant. (7) The remaining processed syslog entries that do not contain additional *variable* terms, are mapped into hash-keys (similar to step (4) above). (8) Upon completion of steps (4) and (7), the hash-key codes can be optimized based on their frequency of appearance. The preliminary results of analyzing the syslogs of a production HPC system called Taurus¹ using the proposed approach shows up to 95% reduction in storage capacity [24]. An interactive demonstration of the use of this anonymization approach on a sample syslog is provided online [25].

In the proposed approach, regular expressions are used for the automatic detection of variable terms within syslog entries. Categorization of automatically detected terms into *sensitive* and/or *meaningful* is performed based on the information in Table II. This information is inferred from the policies and conditions of the host high-performance computing system. Automatically detected variable terms which do not belong to any of the *sensitive* and *meaningful* categories are considered as *semantic-less*. A variable length hash algorithm is used to encode the syslog entries. The encoding step is described in greater details at the end of this section.

Table I contains fifteen main regular expressions (out of thirty-eight) which are used to detect variable terms in syslog entries. The order of their application is significant since certain patterns are subsets of other patterns. Even though most variables can be detected with these regular expressions, in an unlikely case of similarity between variables and constants, the regular expression may not be able to differentiate between constants and variables correctly. For example the username *panic* may be misinterpreted as a constant value like *kernel panic*. In such scenarios, the undetected variables are considered as constants (or vice versa) and appear as a new event pattern. Encoding event patterns in the final step of the proposed approach guarantees the highest attainable level of anonymization.

As the first step, the *quality* of syslog entries needs to be quantified. The product of four characteristics of syslog entries defines the syslog entry quality: (1) sensitivity, (2) semantic, (3) length, and (4) usefulness. To render uniform the impact of all characteristics, their significance is normalized in the range of 0 to 1, and the negative parameters are replaced with their reverse positive counterparts. Therefore, the effective parameters are *nonsensitivity*, *semantic*, *reduction*, and *usefulness*. The *nonsensitivity* parameter of a syslog entry can take any value in the range of 0 (most sensitive) to 1 (most nonsensitive) denoting the best value. The parameter *semantic* can also take any value in the range of 0 (least semantic) to 1 (highest semantic), with 1 representing a highly relevant syslog entry.

The length of syslog entries can be interpreted as the size of syslog entry. The *reduction* of syslog entries size may also take any value in the range of 0 (no reduction) to 1 (most reduction). Size *reduction* can be achieved via any

¹<https://doc.zih.tu-dresden.de/hpc-wiki/bin/view/Compendium/SystemTaurus>

TABLE I

THE VARIABLE TERMS IN THE TAURUS SYSLOGS CAN BE DETECTED WITH THIRTY-EIGHT REGULAR EXPRESSIONS. OUT OF THOSE, THE MAIN FIFTEEN MACHINE-INDEPENDENT REGULAR EXPRESSIONS ARE SHOWN HERE. THESE CAN BE USED TO IDENTIFY THE VARIABLE TERMS WITHIN SYSLOGS FROM ANY COMPUTING SYSTEM.

Variable type	Regular expression
Path	(([\\(\\s\\,\\>\\: =)]([\\/] [a-z0-9_\\.\\-\\:]*))+
Version	(([\\w\\.\\-]+x86_64)
Email	([a-z0-9_\\.\\-]+@[a-z0-9_\\.\\-]+[a-z]+)
DateTime	(\\d{4}-\\d{2}-\\d{2})T(\\d{2}:\\d{2}:\\d{2})
IPv4	(\\d+\\.\\d+\\.\\d+\\.\\d+)
Port	(([\\W]) (port \\d+)
Parameter	(\\\$ [a-z0-9_]+)
URID	(uid=[\\w\\-]+)
User	(for) ((user\\) * [a-z0-9_]+)
Library	([a-z0-9_\\.\\-]+\\.so(\\.\\d+)*)
Hardware address	(0[x] [a-f0-9]+\\+0[x] [a-f0-9]+)
Hex Number	(0[x] [a-f0-9]+)
Percentage	(\\d+\\. * [\\d]*%)
Serial number	((\\s) ([a-f0-9_\\.\\-]+\\:)+ (\\s))
Size	((^a-z0-9]) (\\d+[bkmg]) ((^a-z0-9])

general lossy or lossless compression algorithm. When the applied compression method does not change the semantic and sensitivity of syslog entries, it is considered as lossless (from the perspective of this work). If the chosen compression method modifies the semantic or sensitivity of syslog entries, in the context of this work, it is taken as an additional level of anonymization rather than compression. A careful consideration of various effective compression algorithms, including Brotli, Deflate, Zopfli, LZMA, LZHAM, and Bzip2, revealed that in affordable time, compression could reduce the data size to 25% of its original size. Therefore the *reduction* of syslogs ranges between 0.75 to 1, where 1 indicates 100% compression and is practically impossible to reach. Every time that a compressed syslog entry is processed, the decompression process imposes an additional performance penalty on the host system. Therefore, the proposed approach in this work uses an encoding algorithm instead of compression algorithms which demand a decompression before accessing the compressed data. The encoded data can be accessed and used without pre-processing.

Unlike the previous three parameters, the fourth parameter, *usefulness*, is boolean and takes 0 or 1 as values. The value of 0 or 1 for *usefulness* denotes that a syslog entry in its current form cannot or can be used for a specific type of analysis, respectively.

$$Q_E = U_E * (n * N_E) * (s * S_E) * (r * R_E) \quad (1)$$

Equation (1) quantifies the *quality* (Q_E) of a syslog entry E as a product of its nonsensitivity (N_E), semantic (S_E), reduction (R_E), and usability (U_E). The coefficients, n , s , and r indicate the importance of nonsensitivity, semantic, and reduction for a specific computing system. The default value for n , s , and r is 1. The value of 0 for usability results in a 0-quality syslog entry and disqualifies the current syslog entry from further analysis. As explained earlier, regardless of system conditions and policies, a *reduction* rate of 75% is

always achievable [26]–[29]. Therefore, the *quality* of a raw syslog entry is calculated using Equation (2).

$$Q_E = 1 * (1 * N_E) * (1 * S_E) * (1 * 0.75) \quad (2)$$

The sensitivity of each syslog entry term is defined based on the policies set up by the computing system administrators. Assume that Table II indicates the sensitivity and the semantic of syslog entry terms of a computing system. The severity degree of each term's sensitivity varies from 0 to 10. This degree is only used to give priority to the individual anonymization steps. Each syslog entry term can be sensitive (Y) or nonsensitive (N). Therefore, in this section, only the boolean sensitivity indicator (Y/N) is considered to denote the sensitivity of each syslog entry term. The same assumptions hold for the semantic of each syslog entry term. Accordingly, each term can be *with or without semantic*. The semantic of each term can be judged from 3 sources. (1) Every sensitive term is also semantic. (2) Every semantic term is marked with "Y" in the semantic table (Table II), (3) All terms not included in the semantic table nor marked with "Y" therein simply have length, are nonsensitive, and have no semantic.

Table III indicates the sensitivity and semantic of each term from the message part of the sample syslog entry based on information from Table II.

TABLE II
CLASSIFICATION OF SYSLOG ENTRY TERMS INTO SENSITIVE AND/OR SEMANTIC. SEVERITY DENOTES THE IMPORTANCE OF THE CHARACTERISTICS FOR THE RESPECTIVE TERMS.

Term	Sensitivity	Severity	Term	Semantic	Severity
User Name	Y	10	accept*	Y	07
IP Address	Y	08	reject*	Y	10
Port Number	Y	01	close*	Y	08
Node Name	Y	03	*connect*	Y	09
Node ID	Y	03	start*	Y	02
Public Key	Y	10	*key*	Y	01
App Name	N	00	session	Y	07
Path / URL	N	00	user*	Y	05

TABLE III
A SAMPLE SYSLOG ENTRY. SENSITIVE AND MEANINGFUL TERMS ARE MARKED WITH "Y" IN THE RESPECTIVE ROWS.

Message	Accepted	publickey	for	Siavash	from	4.3.2.1
Sensitive	-	-	-	Y	-	Y
Semantic	Y	Y	-	Y	-	Y

The nonsensitivity (N_E) of a syslog entry E is defined as $\frac{\text{Number of nonsensitive terms in entry } E}{\text{Total number of terms in entry } E}$. The semantic (S_E) of a syslog entry E is defined as $\frac{\text{Number of terms with semantic in entry } E}{\text{Total number of terms in entry } E}$. Calculating these properties for the sample syslog entry E_1 from Table III, with the information from Table II, results in: $N_{E_1} = \frac{4}{6}$ and $S_{E_1} = \frac{4}{6}$, respectively. The quality of the sample syslog (Q_{E_1}) is then obtained with Equation (2) to be $Q_{E_1} = 1 * \frac{4}{6} * \frac{4}{6} * 0.75 \approx 0.33$. The steps for performing a full anonymization with the proposed approach on the sample syslog entry E_1 from Table III are shown in Table IV.

The encoding algorithm used in this work is the variable length hash algorithm SHAKE-128 [30],

TABLE IV
ANONYMIZATION AND ENCODING (HASHING) OF THE SAMPLE SYSLOG
ENTRY FROM TABLE III.

	Message	Accepted	publickey	for	Siavash	from	4.3.2.1
A)	Sensitive	-	-	-	Y	-	Y
	Semantic	Y	Y	-	Y	-	Y

$$Q_{E_1} = 1 * 0.67 * 0.67 * 0.75 \approx 0.33$$

	Anon. #1	Accepted	publickey	for	#USR#	from	4.3.2.1
B)	Sensitive	-	-	-	-	-	Y
	Semantic	Y	Y	-	-	-	Y

$$Q_{E_1} = 1 * 0.83 * 0.50 * 0.75 \approx 0.31$$

	Anon. #2	Accepted	publickey	for	#USR#	from	#IP4#
C)	Sensitive	-	-	-	-	-	-
	Semantic	Y	Y	-	-	-	-

$$Q_{E_1} = 1 * 1.00 * 0.33 * 0.75 \approx 0.25$$

	Anon. #3	Accepted	#KEY#	for	#USR#	from	#IP4#
D)	Sensitive	-	-	-	-	-	-
	Semantic	Y	-	-	-	-	-

$$Q_{E_1} = 1 * 1.00 * 0.17 * 0.75 \approx 0.125$$

	SHAKE-128	caa5002d					
E)	Sensitive	-	-	-	-	-	-
	Semantic	Y	-	-	-	-	-

$$Q_{E_1} = 1 * 1.00 * 1.00 * 0.81 \approx 0.81$$

[31] with 32-bit output length adjustable based on the system requirements. All syslog entries that follow the pattern of the sample syslog entry E_1 : "Accepted publickey for Siavash from 4.3.2.1", regardless of the values which they carry, after 'constantification' are identical to: "Accepted publickey for #USR# from #IP4#". **This string is an event pattern.** Event patterns are constant strings with a certain semantic meaning. Replacing them with a shorter identifier does not change their meaning, as long as the identifier replacing a particular event pattern is known. Therefore, in this work, a hashing function is used to transform event patterns from syslogs to shorter single-term identifiers. Using a hashing function guarantees that an event pattern is always converted to an identical identifier (hash-key). The identifier (hash-key) carries the same semantic as the event pattern, in an 8-character string. The identifier "caa5002d" in comparison with the original string of "Accepted publickey for Siavash from 4.3.2.1" with 43 characters, represents an 81% decrease in the string length. Apart from shortening the syslog entries, using identifiers also reduces the number of terms in each syslog entry, which in turn, results in significant performance improvement of further processing of syslog entries.

IV. METHODOLOGY

We selected a thirteen-month collection of syslogs between February 01, 2016 and February 28, 2017, from the Taurus

production HPC cluster as the source of information in the present study. Taurus is a Linux-based parallel cluster with 2014 computing nodes. It employs Slurm [32] as its batch system. Taurus' 2014 computing nodes are divided into six islands, mainly based on their processing units type: CPUs (Intel's Haswell, Sandy Bridge, Triton, Westmere), and GPUs.

The thirteen-month collection includes syslog entries from all 2014 Taurus computing nodes. The syslog daemons on the computing nodes are configured to submit syslog entries to a central node. The central node, in turn, sends the entries to a syslog storage node. On this storage node, daily syslog entries are accumulated according to their origin into different *log files*. Therefore, the thirteen-month collection includes 2014 system log files per day (one log file for each computing node).

The number of syslog entries generated by a computing node per day depends on various factors, including system updates and node failures. For the thirteen-month period of this study, approximately 984.26GiB of syslogs were collected, which comprise 8.6 billion syslog entries. Fig. 3 illustrates the distribution of syslog entries among the first 100 nodes of each island. A row in an island indicates a node and a column represents a full month period between February 2016 and March 2017.

Various causes, such as scheduled maintenance or node failures, are responsible for a certain percentage of errors during the collection of syslog entries. The *completeness* of the syslog collection process can be measured by considering the presence of a *log file* as the indicator of the gathering of syslog entries from a particular node on a given day. Based on this definition, the syslog collection *completeness* for the specific time interval in this work is 97%. The red lines in Fig. 4 indicate the 3% of *missed* (uncollected) syslogs.

Most of the missing 3% syslog entries have been lost over the course of three days, marked at the top of Fig. 4 with letters A, B, and C. The reasons for their occurrence was (A) scheduled maintenance, (B) reaction of automatic overheating protection mechanism, and (C) failure of the central syslog collection node.

V. RESULTS

The proposed anonymization approach has been applied to a thirteen-month collection of Taurus syslog entries. During this process, the sensitivity and semantic of each term needed to be identified based upon the policies of Taurus HPC cluster. According to the user privacy and data protection act of the *Center for Information Services and High Performance Computing*, at the Technical University of Dresden (TUD), Germany, HPC system usage information may be anonymously collected from the users and shared with research partners. This information includes, yet is not limited to, various metrics about processors, networks, storage systems, and power supplies [33]. Other types of information are processed according to the *IT* [34] and *identity management* [35] regulations of TUD. Based on these regulations, certain data are considered sensitive and must, therefore, remain confidential [36]. To the best of our knowledge, the information in Table V captures

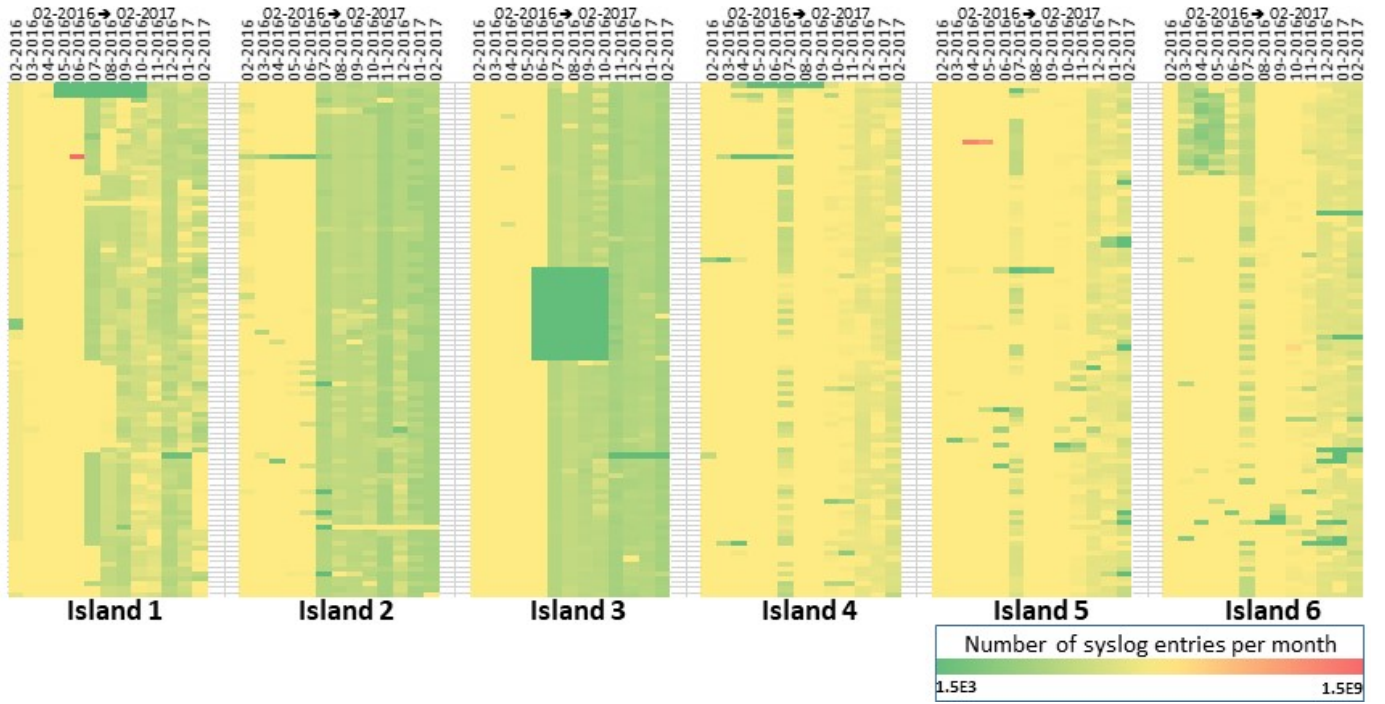


Fig. 3. The number of syslog entries per node, per month. In each island, a column represents (from left to right) a month between February 2016 and March 2017. A row denotes a computing node in each island. The intersection between rows (nodes) and columns (months) is called a cell. The heat color within a cell illustrates the relative number of syslog entries for a particular node in a specific month.

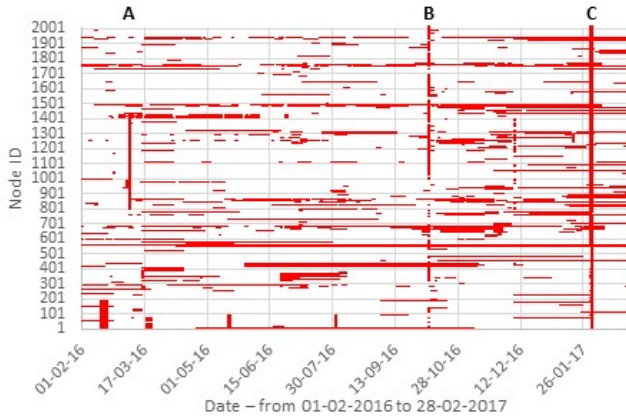


Fig. 4. Illustration of the syslog entries collection gaps. Approximately 3% of the syslog entries collected over thirteen months were not correctly recorded. The data loss occurred at three distinct points in time, identified as three vertical lines. The causes of this data loss are (A) scheduled maintenance, (B) reaction of automatic overheating protection mechanism, and (C) failure of the central syslog collection node.

TABLE V
SYSLOG ENTRY SENSITIVITY ACCORDING TO THE TUD PRIVACY REGULATIONS

Term	Sensitivity	Severity
Surname	Y	10
Firstname	Y	10
Title	Y	10
User type (employee, student, guest)	Y	10
User name	Y	10
Password	Y	10
Login status (active, disabled)	Y	10
User ID (identification of Unix users)	Y	10
Home (Path to home directory)	Y	10
Shell (default shell)	Y	10
Group ID (belonging to Unix groups)	Y	10
Mail addresses (TUD addresses)	Y	10
IP Address	Y	08
Port Number	N	00
Node Name	N	00
Node ID	N	00
Public Key	Y	08
App Name	N	00
Path / URL	Y	01

the data sensitivity according to the TUD privacy regulation in force. From this Table V, one can note that certain syslog entry terms, such as node names or port numbers, may remain unchanged. The use of the proposed approach on 8.6 billion syslog entries from Taurus of an uncompressed size of 985 GiB, according to the TUD privacy regulations revealed seven facts.

(1) Only approximately 35% of the syslog terms are sensitive and need to be anonymized. Therefore approximately 65%

of terms remained untouched (Fig. 5).

(2) The anonymization of sensitive terms has less than 0.5% impact on syslog size reduction.

(3) The quality of most entries degraded post-anonymization.

(4) Approximately 2,000 unique event patterns were discovered.

(5) More than 90% of syslog entries are based on 40 event patterns (hereafter *frequent patterns*). All other non-frequent

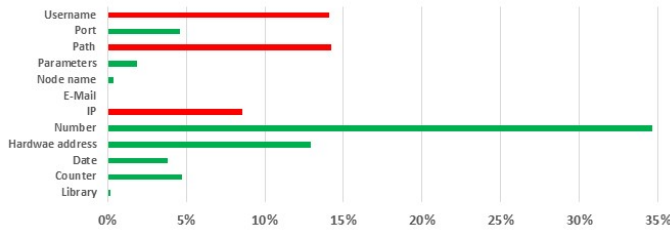


Fig. 5. Percentage of sensitive and nonsensitive terms within the thirteen-month-long collection of syslogs. The red bars indicate the percentage of sensitive terms while the green bars indicate the percentage of nonsensitive terms. The sensitive terms sum up to approximately 35% of all terms in the collection.

event patterns together are responsible for less than 10% of syslog entries. For instance, more than 15% of the syslog entries have the `(#USER#) cmd (#PATH#)` pattern.

(6) A small percentage of syslog entries (approximately 5%) among the non-frequent event patterns do not contain any variable terms in their original form (e.g., `disabling lock debugging due to kernel taint`).

(7) According to the current TUD privacy regulations, almost all syslog entries lose their added semantic after anonymization (e.g., `"failed password for #USER# from #IPv4# port 32134 ssh2"`).

The only remaining useful information in these cases is the semantic of the event pattern itself. For the above example, the useful semantic is that `authentication via ssh failed`. Based on the above seven observations about syslog entries on Taurus, we can state that: (1) The anonymized syslogs consist of approximately 90% semantic-less entries (after mandatory anonymization), (2) Approximately 5% of the entries are constant, that is they do not have any variable terms, and approximately 5% are entries with semantic (and retained their useful properties even after anonymization). Following the necessary anonymization, (90 + 5)% of syslog entries no longer have semantic and can be converted to hash-keys. The 5% of syslog entries which carry added semantic even after anonymization, should remain untouched.

The Table VI illustrates a sample of four syslog entries in three different stages of anonymization. The Table VII is a reference to the meaning of each of the hash keys. Together with the anonymized syslogs and according to the privacy regulations, the information in Table VII may also be fully/partially published.

The data in part (B) of Table VI follow the main anonymization guidelines. This fact enables their inclusion in the present work. However, since syslog entries lengths have been reduced, the data in part (C) of Table VI delivers the very same semantic as part (B), at a much smaller length.

The usefulness of the anonymized and hashed information from the thirteen-month syslog collection remains identical. Reprocessing the results from an earlier work [37], in which the correlation of failures in Taurus was analyzed, via the new anonymization and encoding approach led to identical

TABLE VI
ANONYMIZATION AND HASHING OF FOUR SYSLOG ENTRIES

(A) Before anonymization

```
1 (siavash) cmd (/home/siavash/config.sh > output.stat)
2 pam_unix(sshd:session): session closed for siavash
3 disabling lock debugging due to kernel taint
4 ACPI: LAPIC (acpi_id[0x55] lapic_id[0xff] disabled)
```

(B) After anonymization, before hashing

```
1 (#USER#) cmd (#PATH# > output.stat)
2 pam_unix(sshd:session): session closed for #USER#
3 disabling lock debugging due to kernel taint
4 ACPI: LAPIC (acpi_id[0x55] lapic_id[0xff] disabled)
```

(C) After anonymization and hashing

```
1 1808e388
2 0964de42
3 59f2da35
4 ACPI: LAPIC (acpi_id[0x55] lapic_id[0xff] disabled)
```

TABLE VII
HASH-KEY REFERENCE TABLE

Hash-key	Meaning
1808e388	A command executed by user
0964de42	A user logged out
59f2da35	Disabling lock debugging due to kernel taint

outcomes. Moreover, due to single-term syslog entries, the processing time was approximately 25% shorter than before.

VI. CONCLUSION

System logs have widely been used in various domains, from system monitoring and performance analysis to failure prediction of different system components. Even though system logs are mainly system dependent, having knowledge about various computing systems improves the general understanding of computing systems behavior. However, due to the vast amount of personal data among the system log entries, users privacy concern impedes the free circulation and publication of system logs. In this work, we examined the trade-off between sensitivity and semantic of system logs. Since after a certain level of anonymization the semantic of system logs may be lost, keeping the semantic-less data is not the best practice.

This work introduced *quality*, the system logs utility function, as a measurable parameter calculated based on nonsensitivity, semantic, reduction, and usefulness of system logs. The goal is to maintain the quality of system log, by pushing all effective parameters to their possible limit. This proposed approach has been applied on a thirteen-month collection of Taurus HPC cluster system logs, between from February 01, 2016 and February 28, 2017. The proposed anonymization approach can guarantee full anonymization of syslog entries via the final encoding step. Apart from the highest degree of anonymization, a total reduction of more than 50% in system log size as well as 25% performance improvement in system log analysis is achievable.

The current hashing function produces larger hash-keys than required to avoid hash-key collisions. Fine tuning of the

hashing function according to the computing system requirements, together with improving the variable term detection, are planned as future work.

ACKNOWLEDGEMENT

This work is in part supported by the German Research Foundation (DFG) in the Cluster of Excellence “Center for Advancing Electronics Dresden” (cfaed). The authors also thank Holger Mickler and the administration team of Technical University of Dresden, Germany for their support in collecting the monitoring information on the Taurus high performance computing cluster.

DISCLAIMER

References to legal excerpts and regulations in this work are provided only to clarify the proposed approach and to enhance explanation. In no event will authors of this work be liable for any incidental, indirect, consequential, or special damages of any kind, based on the information in these references.

REFERENCES

- [1] T. C. Redman, *Data Driven: Profiting from Your Most Important Business Asset*. Harvard Business Press, 2008.
- [2] “European Commission Decision,” <http://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:32000D0520>, [Online; accessed 06-June-2017].
- [3] “General data protection regulation,” <http://gdpr-info.eu/art-4-gdpr/>, [Online; accessed 06-June-2017].
- [4] L. Sweeney, “Simple demographics often identify people uniquely,” *Carnegie Mellon University, Data Privacy*, 2000, working paper.
- [5] R. Dahlberg and T. Pulls, “Standardized Syslog Processing: Revisiting Secure Reliable Data Transfer and Message Compression,” Karlstad, Sweden, 2016.
- [6] “New rsyslog 7.4.0,” <http://www.rsyslog.com/7-4-0-the-new-stable/>, [Online; accessed 06-June-2017].
- [7] “Logstash, centralize, transform and stash your data,” <http://www.elastic.co/products/logstash>, [Online; accessed 06-June-2017].
- [8] S. Sanjappa and M. Ahmed, “Analysis of logs by using logstash,” in *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*. Springer Singapore, 2017, pp. 579–585.
- [9] “Loggly, log management,” <http://www.loggly.com/>, [Online; accessed 06-June-2017].
- [10] “Siem, log management, compliance,” <http://www.logsign.com/>, [Online; accessed 06-June-2017].
- [11] “Blazing-fast log management and server monitoring,” <http://www.scalyr.com>, [Online; accessed 06-June-2017].
- [12] A. Gholami, E. Laure, P. Somogyi, O. Spjuth, S. Niazi, and J. Dowling, “Privacy-preservation for publishing sample availability data with personal identifiers,” *Journal of Medical and Bioengineering*, vol. 4, no. 2, 2015.
- [13] M. Templ, A. Kowarik, and B. Meindl, “Statistical disclosure control methods for anonymization of microdata and risk estimation,” <http://cran.r-project.org/web/packages/sdcMicro/index.html>, [Online; accessed 06-June-2017].
- [14] C. Dai, G. Ghinita, E. Bertino, J.-W. Byun, and N. Li, “TIAMAT: A Tool for Interactive Analysis of Microdata Anonymization Techniques,” *Proc. VLDB Endow.*, vol. 2, no. 2, pp. 1618–1621, 2009.
- [15] M. Ciglic, J. Eder, and C. Koncilia, “Anonymization of data sets with null values,” *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXIV: Special Issue on Database- and Expert-Systems Applications*, pp. 193–220, 2016.
- [16] “UTD anonymization toolbox,” <http://cs.utdallas.edu/dspl/cgi-bin/toolbox>, [Online; accessed 06-June-2017].
- [17] X. Xiao, G. Wang, and J. Gehrke, “Interactive anonymization of sensitive data,” in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD ’09. New York, NY, USA: ACM, 2009, pp. 1051–1054.
- [18] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian, “L-diversity: privacy beyond k-anonymity,” in *22nd International Conference on Data Engineering (ICDE’06)*, April 2006, pp. 24–24.
- [19] A. Meyerson and R. Williams, “On the complexity of optimal k-anonymity,” in *Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ser. PODS ’04. New York, NY, USA: ACM, 2004, pp. 223–228.
- [20] R. J. Bayardo and R. Agrawal, “Data privacy through optimal k-anonymization,” in *21st International Conference on Data Engineering (ICDE’05)*, April 2005, pp. 217–228.
- [21] C. Rath, “Usable privacy-aware logging for unstructured log entries,” in *11th International Conference on Availability, Reliability and Security (ARES)*, Aug 2016, pp. 272–277.
- [22] “Privacy-aware logging made easy,” <http://github.com/nobecutan/privacy-aware-logging>, [Online; accessed 06-June-2017].
- [23] “The syslog protocol,” <http://tools.ietf.org/html/rfc5424>, [Online; accessed 06-June-2017].
- [24] S. Ghiasvand and F. M. Ciorba, “Toward Resilience in HPC: A Prototype to Analyze and Predict System Behavior,” Poster at International Supercomputing Conference (ISC), June 2017.
- [25] “Demonstration of anonymization and event pattern detection,” <http://www.ghiasvand.net/u/hpcmaspa17>, [Online; accessed 06-June-2017].
- [26] J. Alakuijala, E. Kliuchnikov, Z. Szabadka, and L. Vandevenne, “Comparison of brotli, deflate, zopfli, lzma, lzham and bzip2 compression algorithms,” <http://cran.r-project.org/web/packages/brotli/vignettes/brotli-2015-09-22.pdf>, [Online; accessed 06-June-2017].
- [27] L. Collin, “A quick benchmark: Gzip vs. bzip2 vs. lzma,” <http://tukaani.org/lzma/benchmarks.html>, [Online; accessed 06-June-2017].
- [28] “Quick benchmark: Gzip vs bzip2 vs lzma vs xz vs lz4 vs lzo,” <http://www.ghiasvand.net/u/compression>, [Online; accessed 06-June-2017].
- [29] M. Mahoney, “10 gb compression benchmark,” <http://mattmahoney.net/dc/10gb.html>, [Online; accessed 06-June-2017].
- [30] G. Bertoni, J. Daemen, M. Peeters, and G. Van Assche, “The KECCAK SHA-3 submission,” <http://keccak.noekoon.org/Keccak-submission-3.pdf>, [Online; accessed 06-June-2017].
- [31] S. Fluhrer, “Comments on fips-202,” http://csrc.nist.gov/groups/ST/hash/sha-3/documents/fips202_comments/Fluhrer_Comments_Draft_FIPS_202.pdf, [Online; accessed 06-June-2017].
- [32] A. B. Yoo, M. A. Jette, and M. Grondona, “SLURM: Simple Linux Utility for Resource Management,” in *Proceedings of 9th International Workshop on Job Scheduling Strategies for Parallel Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 44–60.
- [33] “Terms of use of the HPC systems at the ZIH, Technical University Dresden, Germany,” http://doc.zih.tu-dresden.de/hpc-wiki/pub/Compendium/TermsOfUse/HPC-Nutzungsbedingungen_20160901.pdf, [Online; accessed 06-June-2017].
- [34] “Order for the Information Technology Facilities and Services and for the Information Security of the Technical University of Dresden (IT-Regulations), Germany,” <http://www.verw.tu-dresden.de/ambek/PDF-Dateien/2016-12/sonstO05.01.2016.pdf>, [Online; accessed 06-June-2017].
- [35] “Order for the Establishment and Operation of an Identity Management System at the Technical University of Dresden, Germany,” <http://www.verw.tu-dresden.de/AmtBek/PDF-Dateien/2011-05/sonstO26.07.2011.pdf>, [Online; accessed 06-June-2017].
- [36] “Information leaflet on IT resources, Technical University Dresden, Germany,” http://tu-dresden.de/zih/dienste/service-katalog/zugangsvoraussetzung/merkblatt?set_language=en, [Online; accessed 06-June-2017].
- [37] S. Ghiasvand, F. M. Ciorba, R. Tschüter, and W. E. Nagel, “Analysis of Node Failures in High Performance Computers Based on System Logs,” Poster at International Conference for High Performance Computing, Networking, Storage and Analysis (SC15), 2015.