**SANDIA REPORT**

SAND2010-2179
Unlimited Release
Printed April 2010

# Statistical Language Analysis for Automatic Exfiltration Event Detection

David G. Robinson

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin company, for the U.S. Department of Energys National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

**NOTICE:** This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2010-2179
Unlimited Release
Printed April 2010

# Statistical Language Analysis for Automatic Exfiltration Event Detection

David G. Robinson
Computer Science and Informatics
Sandia National Laboratories
Albuquerque, NM 87185
drobin@sandia.gov

**Abstract**

This paper discusses the recent development a statistical approach for the automatic identification of anomalous network activity that is characteristic of exfiltration events. This approach is based on the language processing method referred to as latent dirichlet allocation (LDA). Cyber security experts currently depend heavily on a rule-based framework for initial detection of suspect network events. The application of the rule set typically results in an extensive list of suspect network events that are then further explored manually for suspicious activity. The ability to identify anomalous network events is heavily dependent on the experience of the security personnel wading through the network log. Limitations of this approach are clear: rule-based systems only apply to exfiltration behavior that has previously been observed, and experienced cyber security personnel are rare commodities. Since the new methodology is not a discrete rule-based approach, it is more difficult for an insider to disguise the exfiltration events. A further benefit is that the methodology provides a risk-based approach that can be implemented in a continuous, dynamic or evolutionary fashion. This permits suspect network activity to be identified early with a quantifiable risk associated with decision making when responding to suspicious activity.

# Contents

**1 Background** **9**

1.1 Cyber Security Problem . . . . . 9

1.2 Notation . . . . . 10

**2 Approach** **11**

2.1 Latent Dirichlet Allocation . . . . . 11

2.2 Document Similarity Measures . . . . . 14

**3 Exfiltration Detection** **17**

3.1 IEEE VAST 2009 Challenge Data . . . . . 19

3.2 Analysis . . . . . 20

**4 Results** **27**

4.1 Discussion . . . . . 28

# List of Figures


2.1 Bag of Words and the Generative Process . . . . . 12

2.2 The Probabilistic Process . . . . . 13

2.3 Implied Bayesian Hierarchical Structure . . . . . 14

2.4 Geometric Interpretation of $\Theta$ . . . . . 15

3.1 High Level Network Map . . . . . 18

3.2 Resulting 50 Clusters . . . . . 21

3.3 Heat Map of Hellinger Matrix - 1000 samples . . . . . 23

3.4 Hellinger Matrix Dendrogram - 500 samples . . . . . 24

3.5 Heat Map of Hellinger Distance - 500 samples . . . . . 25

# List of Tables

3.1 Sample of Raw Network Event Data . . . . . 19

3.2 Sample of Tokenized Event Data . . . . . 20

4.1 Final Set of Suspect Network Events . . . . . 27

# Chapter 1

# Background

Information is a collection of data: documents, images, symbols, network files, etc. The field of information analysis is broadly concerned with the problem of organizing this data to facilitate the discovery and understanding of the structure associated with the data. When the information is a collection of documents, the goals are generally to construct a framework to assist in identifying critical elements or concepts for further action and organize the information in the corpus for retrieval by users. Bits of information that present a challenge to an organization (i.e., anomalies) often provide unique opportunities for discovery.

Relatively new to information analysis, the mathematical techniques associated with probabilistic latent semantic analysis are becoming an increasingly powerful analysis suite for organizing information. This family of methods includes, for example, Pachinko Allocation Models (PAM) [5] and Latent Dirichlet Allocation (LDA) [2].

## 1.1 Cyber Security Problem

From an introspective point of view, of particular concern to an organization is the leaking of critical information via network transfer. Identification of data exfiltration is particularly challenging since organizations rely on a broad range of network communication methods. It is often not possible to determine if a leak has even occurred. Exfiltration can take many forms: insider information transfer, Trojan backdoor traffic, or a system administrator leaving file access unprotected. There is intense focus by cyber security experts on malicious transfer of information by an insider.

Cyber security experts depend heavily on a rule-based framework for initial detection of suspect network events. The application of the rule set typically results in an extensive list of suspect network events that are then further explored manually for suspicious activity. The ability to identify anomalous network events is heavily dependent on the experience of the security personnel wading through the network log. Limitations of this approach are clear: rule-based systems only apply to exfiltration behavior that has previously been observed, and experienced cyber security personnel are rare commodities.

We have developed an alternative approach to automatically identify unusual or anomalous network traffic generated within an organization. The method involves the analysis of external

network header information through a fast LDA algorithm. Unusual patterns in network events are successively identified and filtered. The result is a smaller, more manageable set of events that can be further resolved by cyber security staff.

Since the new methodology is not a discrete, rule-based approach, it is more difficult for an insider to disguise the exfiltration events. A further benefit is that the methodology provides a risk-based approach that can be implemented in a continuous, dynamic or evolutionary fashion. This permits suspect network activity to be identified early with a quantifiable risk associated with decision making.

Following a brief review of notation, the fundamental statistical concepts associated with Latent Dirichlet Allocation (LDA) will be briefly introduced. Following that will be a brief introduction to the various information measures used in the *analysis*. Finally, the structure of the network data to be analyzed will be introduced, and the results of the analysis will be summarized.

## 1.2 Notation

- $i = 1, \ldots, M$ = number of documents
- $j = 1, \ldots, N_i$ = number of words in document $i$
- $k = 1, \ldots, T$ = number of topics

- $\mathbb{W} = \{w_1, w_2, \ldots, w_M\}$ = set of all words ($\sum w_M = V =$ vocabulary)
- $\mathbb{Z} = \{z_1, z_2, \ldots, z_T\}$ = set of all topics

- $\phi_k$ = probability density function of words in a topic $k$, e.g. $\phi_1 = \{\phi_{11}, \ldots, \phi_{1T}\}$
- $\theta_i$ = probability density function of topics in document $i$, e.g. $\theta_1 = \{\theta_{11}, \ldots, \theta_{1T}\}$
- $\Theta = \{\theta_1, \ldots, \theta_M\}$
- $w_{ij}$ = word $j$ in document $i$
- $z_{ij}$ = topic for word $j$ in document $i$

# Chapter 2

# Approach

## 2.1 Latent Dirichlet Allocation

Fundamental to our analysis are the concepts associated with natural language processing. A common approach in text analysis is Latent Semantic Analysis (LSA) [3]. LSA is a popular geometric method for characterizing the relationships between documents and the words within the documents. A document is considered to be composed of a collection of words: a *bag of words*, where word order or grammar are not considered important. Attempts to statistically interpret results from LSA have been misleading. Overlaying a statistical framework on LSA relies on the unrealistic assumption that the relationship between documents and words is a Gaussian random process; contrary to what has been observed.

However, as noted by Hofmann [4], if we examine the *bag of words* assumption closer, a possible probability structure is implied. de Finetti's Theorem states that an exchangeable sequence of random variables is a mixture of independent and identically distributed random variables.

A finite set of random variables is said to be exchangeable if the joint distribution is invariant to permutation, e.g. $p(z_1, z_2, \ldots, z_T) = p(z_T, z_1, \ldots, z_2)$. Exchangeable observations are independent conditioned on some non-observable (latent) random variable.

The impact of exchangeability is subtle but to get an appreciation of the importance: frequentists assume that observations are independent and identically distributed while Bayesians assume that the observations are exchangeable. Note that the bag of words assumption is still applicable. The assumption of de Finettis Theorem allows the use of generative models for describing how documents can be constructed and provides additional insight into the data. Order is still not important but the resulting probability structure allows for consideration of the context in which a word is used. An important possibility in natural language analysis.

The assumption of de Finetti's Theorem allows the use of generative models for describing how documents can be constructed:

- Select document $d_i$ with probability $p(d_i)$
  - for each document, pick a topic $z_k$ with probability $p(z_k|d_i, \theta)$
  - select a word $w_j$ from topic $z_k$ with probability $p(w_j|z_k, \phi)$

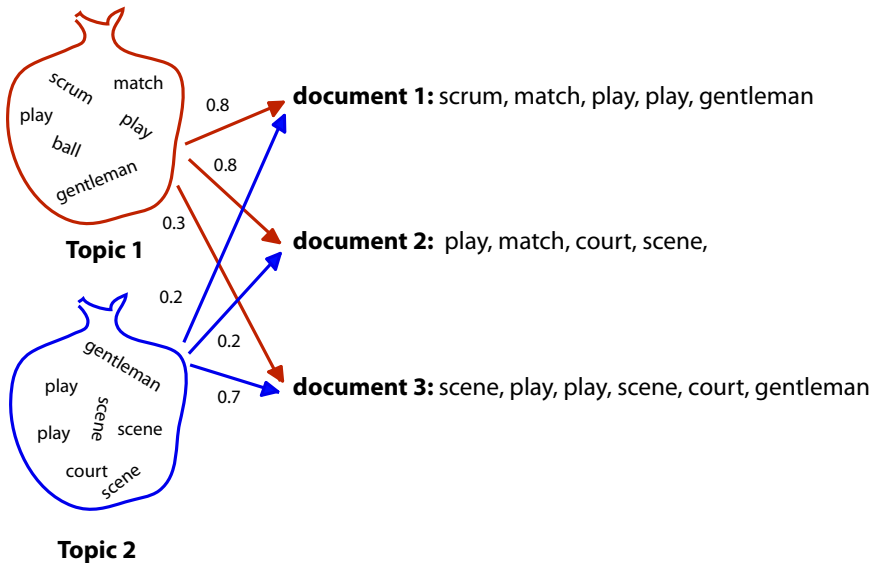


**Figure 2.1.** Bag of Words and the Generative Process

- the result is a document and word pair: $(d, w)$

- Repeat for all words in all documents

The generative process for text analysis, first suggested by Hofmann [4], is depicted in Figure 2.1. Selection of words from the various bags are based on some implied probabilistic mixture of topics and the likelihood of words being chosen is conditioned on the Topic Bag selected. Our goal, however, is depicted in Figure 2.2. We are provided a corpus of documents and we wish to characterize the mixture of topics associated with each document and also characterize the distribution of words in the various Topic Bags.

If we assume a generative process for documents, then our analysis is characterized by a hierarchical Bayesian structure (Figure 2.3) to estimate the relationships between documents and topics, and between topics and words given a corpus $D = (d, w)$ . The goal then is, given the values of $(d, w)$, to evaluate the conditional distributions:

- the word distribution $p(w_j|z_k) = \phi_{jk}$ for each topic k and,

- The topic distribution $p(z_k|d_i) = \theta_{ki}$ for each document i

If we further assume particular distributions: $z_k|\theta \sim Multi(\theta)$, $w_j|\phi \sim Multi(\phi)$, $\theta \sim Dir(\alpha)$, and $\phi \sim Dir(\beta)$, then we have Latent Dirichlet Allocation (LDA). A variety of methods have been proposed for estimating the free parameters $\Theta$ and $\Phi$; our choice is Gibbs sampling since it is stable, relatively fast, and resistant to local minima. The hyperparameters $\alpha$ and $\beta$ can be vector

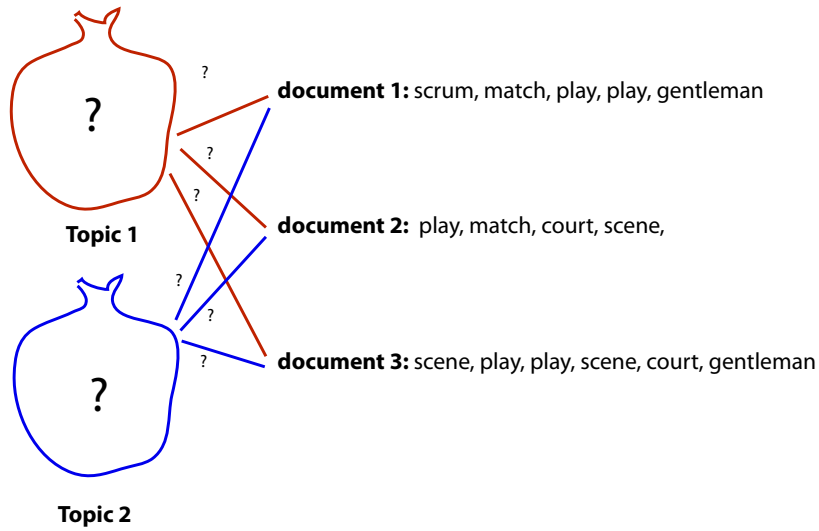


**Figure 2.2.** The Probabilistic Process

valued, but for this discussion we will assume symmetric Dirichlet distributions, i.e. $\alpha = \sum_k \alpha_k / T$ and $\beta = \sum_j \beta_j / V$ will be single valued.

As a side note, the runtime complexity of the Gibbs sampler is $O(MTL)$ where $M$ is the number of documents, $T$ is the number of topics, and $L$ is the average number of words in a document.

It is possible to generate a pseudo-corpus of documents. Define a $\Phi = \left[\phi_{jk}\right]_{T \times V}$ matrix of word parameters such that $\phi_{jk} = p(w_j|z_k)$ is a multinomial distribution describing the probability that word $j$ is associated with topic $k$. Define a $\Theta = \left[\theta_{jk}\right]_{M \times T}$ matrix of topic variables such that $\theta_{jk}|\alpha_k$ is a random variable characterizing that probability that topic $k$ is associated with document $j$ conditioned on the T-length vector of parameters $\alpha_k$. $T$ is the presumed number of topics fixed at the start of the analysis, $\Theta$ and $\Phi$ are estimated during the course of the statistical analysis.

- choose length of document $i$, $N_i \sim Poisson(\eta)$
- Choose topic probability density function for document, $\Theta = (\theta_1, \ldots, \theta_k) \sim Dirichlet(\alpha)$
- For each of the $N$ words
  - choose the topic distribution $k$ for each word: $z_k \sim Multi(\Theta; 1)$
  - choose word $w_j \sim p(w_j|z_k, \beta)$ where $\beta_{jk} = p(w_j|z_k)$

$$p(w, z, \Theta|\alpha, \beta) = p(\Theta|\alpha) \prod_{n=1}^{N} p(z_k|\Theta) p(w_j|z_k, \beta) \tag{2.1}$$

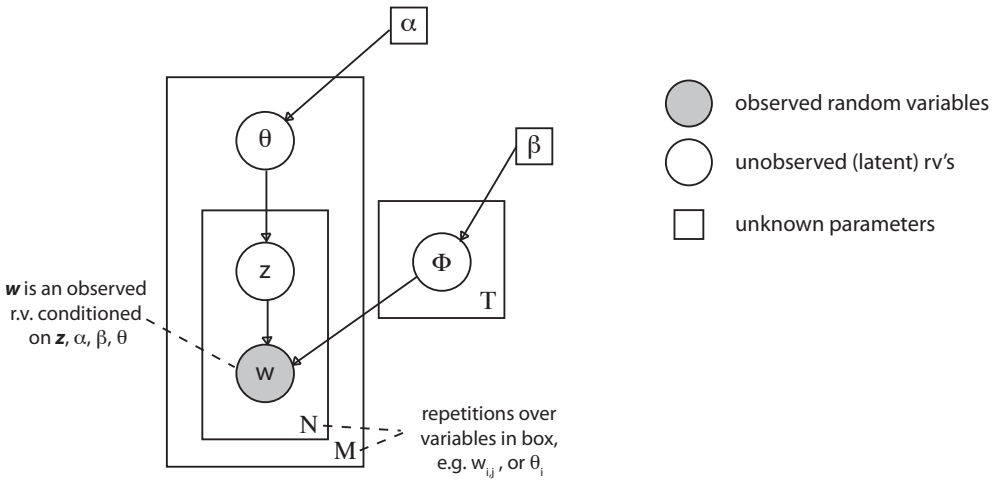


**Figure 2.3.** Implied Bayesian Hierarchical Structure

$$p(w|\alpha,\beta) = \int_{\Theta} p(\Theta|\alpha) \prod_{n=1}^{N} \sum_{z_k=1}^{T} p(z_k|\Theta) p(w_j|z_k,\beta) d\Theta \tag{2.2}$$

$$p(d|\alpha,\beta) = \prod_{i=1}^{M} p(w|\alpha,\beta) \tag{2.3}$$

(major subscripts have been removed to avoid confusion)

## 2.2 Document Similarity Measures

Current topic analysis methods generally focus on characterizing the similarity between documents by estimating the similarity between the topics contained in the documents. The general approach involves comparing the similarity of their topic distributions $\theta_r$ and $\theta_s$. For this reason the $\Theta$ matrix is a key element in measuring similarity. Figure 2.4 depicts a geometric interpretation of $\Theta$. Each document, $d_i$, resides on a convex hull where the location indicates the probability mixture of topics within that document $p(z_k|d_i)$. The documents are constrained by $\sum_k^T p(z_k|d_i) = 1$. The similarity between the documents can be characterized using one of a number of measures. Each of the following were explored for possible application to our particular problem.

**Kullback-Leibler** Common similarity measure for probability distributions is the Kullback-Leibler divergence measure:

$$D(\theta_r,\theta_s) = \sum_{k=1}^{T} \theta_r \log_2 \frac{\theta_r}{\theta_s} \tag{2.4}$$

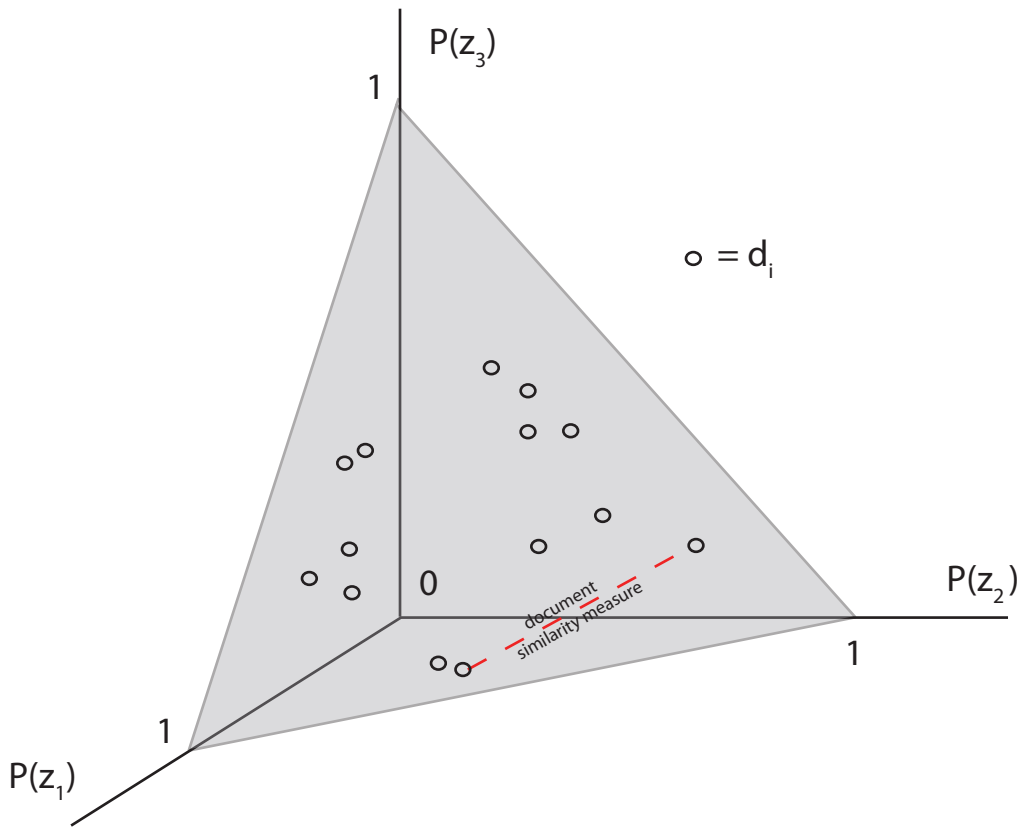


**Figure 2.4.** Geometric Interpretation of Θ

KL is asymmetric and a symmetric alternative is:

$$KL_s(\theta_r, \theta_s) = \frac{1}{2}\left[D(\theta_r, \theta_s) + D(\theta_s, \theta_r)\right] \tag{2.5}$$

**Jensen-Shannon** A closely related similarity measure is the Jensen-Shannon divergence measure:

$$D_{JS}(\theta_r, \theta_s) = KL_s(\theta_r, \theta_q) = \frac{1}{2}\left[D(\theta_r, \theta_q) + D(\theta_s, \theta_q)\right] \tag{2.6}$$

where:

$$\theta_q = (\theta_r + \theta_s)/2 \tag{2.7}$$

The Jensen-Shannon divergence is equal to 1/2 the Jeffrey Divergence.

**Meila Divergence** Meila [6] proposed a similarity measure between document clusters based on information theory. Define the entropy of a document cluster r:

$$H_r = -\sum_{k=1}^{K} \theta_{rk} \log_2 \theta_{rk} \tag{2.8}$$

and:

$$E\left[\theta_r\right] = \sum_m \theta_{rm}/M \tag{2.9}$$

$$E\left[\theta_{rs}\right] = \sum_m \theta_{rm}\theta_{sm}/M \tag{2.10}$$

$$I(r,s) = \sum_{j=1}^{J}\sum_{k=1}^{T} E\left[\theta_{jk}\right] log_2\left[\frac{E\left[\theta_{jk}\right]}{(E\left[\theta_j\right])(E\left[\theta_k\right])}\right] \tag{2.11}$$

then:

$$D_{r,s} = H_r + H_s - 2I(r,s) \tag{2.12}$$

**Hellinger Distance** Hellinger Distance is a measure of the distance between populations with multivariate distributions having probability density functions r and s. The distance is given by $\sqrt{2(1-p)}$ where:

$$p = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} r(\theta_r) s(\theta_s) d\theta_r d\theta_s \tag{2.13}$$

Blie [1] suggests the following form of the Hellinger distance as a measure of the similarity between documents i, j :

$$D_{ij} = E\left[\sum_{k=1}^{K}\left(\sqrt{\theta_{ik}} - \sqrt{\theta_{jk}}\right)^2 | w_i, w_j\right] \tag{2.14}$$

$$D_{ij} = \sum_{k=1}^{T}\left(\sqrt{\theta_{ik}} - \sqrt{\theta_{jk}}\right)^2 / T \tag{2.15}$$

# Chapter 3

# Exfiltration Detection

The level of cyber attack is becoming increasingly more sophisticated and targeted. The attacks are also becoming more surreptitious and persistent. There has been a movement away from individual activity and an increasing number of actions by organized criminal syndicates and nation states. Our focus here is the transfer of sensitive information from an organization computer system to an external agency. Apart from an insider using removable media, exfiltration of data typically involves moving the data from the target system to an external platform through the use of a network channel.

Figure 3.1 depicts a typical computer network for a research organization. Monitoring for exfiltration generally occurs at one or more of the DMZ firewalls. The network traffic at a firewall might be on the order of 100s of thousands network events per minute.

Information exportation may involve an insider or the use of malware to establish a foothold and then the use of the malware to discover additional targets. Once the target has been breached, it is possible for the malware to harvest sensitive data using either automated or manual search methods. Subsequent exfiltration will generally appear benign in a routine log review. Serious review of network logs can be time consuming and require substantial experience to identify anomalous network events associated with exfiltration. When the rate of network events can exceed a thousand events per second, the task of identifying suspicious events borders on the impossible.

The approach to reducing the risk of data exfiltration suggested here is to reduce the number of network events to be reviewed to a manageable level. Contrary to popular opinion, you are not required to fully understand what normal network behavior is to identify abnormal behavior. You just need to be able to identify *less normal* behavior.

Current cyber security methods depend heavily on a long list of discrete rules; and the list is becoming longer leading to potential conflicts. In addition, current methods rely heavily on years of experience by cyber security analysts to identify suspicious network events

However, rules only apply to things that you or others have observed; the proposed approach allows for discovery of unknowns. Gaining experience at identifying the moving target can be costly in time and money. The proposed approach provides a fast, simple method for performing an initial filtering of network events.

To demonstrate the proposed approach and avoid issues with privacy and security associated

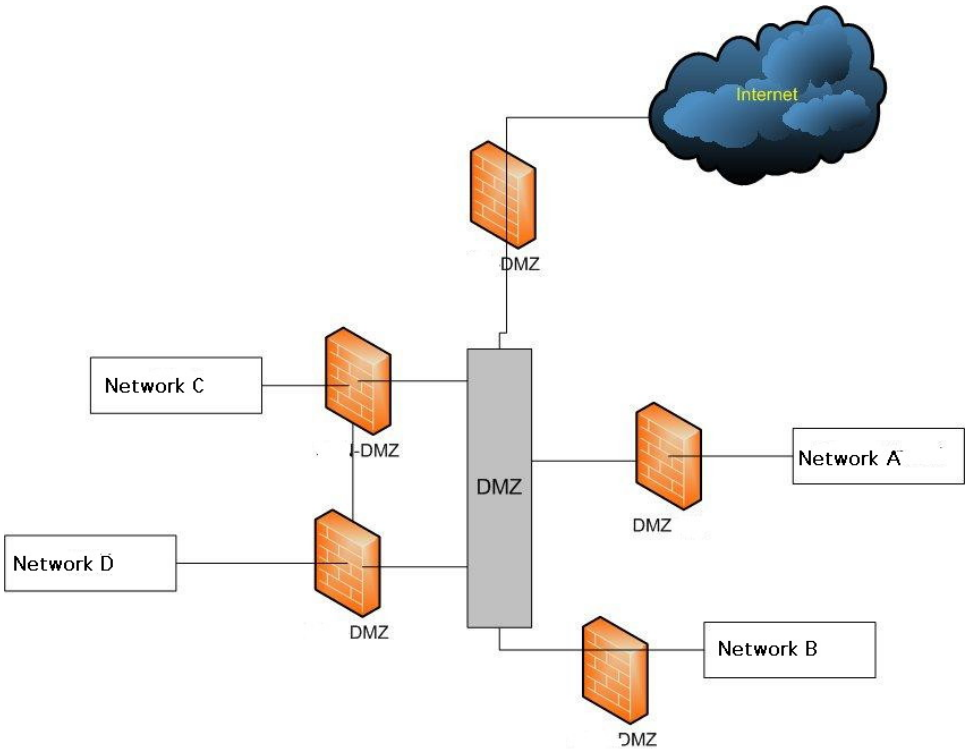


**Figure 3.1.** High Level Network Map

with actual network traffic, a publicly available data set is used in the analysis. The following section discusses the source and nature of the data.

## 3.1 IEEE VAST 2009 Challenge Data

The following problem was posed as part of the IEEE VAST Challenge 2009 [7].

*There are problems at the U.S. Embassy in Flovania, a small land-locked country situated between the nations of Trium to the north and east, Posana to the west, and the Republic of Transak to the southwest. The network security team recently found irregularities while reviewing Embassy network traffic logs.*

*Finding no security issues that could explain the anomalies, they notified the Embassy Counterintelligence Officer. Upon further investigation, the CI officer identified certain espionage You have been requested to help him with the investigation. The 60 Embassy employees have 24x7 access to the facility, with access to the classified information room.*

*A months worth of network traffic logs is available. Each employee has been assigned a desktop computer with a static IP address for use in their daily duties. The network traffic log data consists of the computer IP address, the employee number of the assigned user, outgoing and incoming activity from the computer including destination site, payload (request and response data) and port number.*

Table 3.1 is the first few lines from the network traffic log. The complete data set is available from the IEEE VAST 2009 website [7]. There are a total of 115,414 network events. The data set is known to contain a set of anomalies, but at the start of the Challenge, the specific exfiltration events are unknown.

| Source IP | Access Date/Time | Destination IP | Socket | Req Size | Resp Size |
|---|---|---|---|---|---|
| 37.170.100.38 | 01/01/08 09:40 AM | 37.170.100.200 | 80 | 7063 | 49591 |
| 37.170.100.38 | 01/01/08 09:43 AM | 37.157.76.124 | 80 | 5171 | 434285 |
| ... | ... | ... | ... | ... | ... |

**Table 3.1.** Sample of Raw Network Event Data

The 151, 414 network events were 'tokenized' prior to analysis. The tokenized records associated with Table 3.1 are presented in Table 3.1. This was done primarily to make interpretation of the results more convenient. Note that all punctuation was removed and the source and destination IP addresses were specifically identified. The AccessDate/Time was reduced to an hour indicator (TMx) and Day of the Week (DOWx). The Request Packet Size and the Response Packet Size were reduced to the scale Extra-Extra Small through Extra-Extra Large (XXS, XS, S, M, L, XL, XXL); this was done using a simple (and arbitrary) scaling to reduce the dimension of the analysis. The seven regions were roughly defined by equal frequency counts of observed packet size.

| Source IP | Access Time | Destination IP | Socket | Req Size | Resp Size | Day of Week |
|---|---|---|---|---|---|---|
| SIP3717010038 | TM09 | DIP37170100200 | SKT80 | RQs_S | RSs_S | DOW2 |
| SIP3717010038 | TM09 | DIP3715776124 | SKT80 | RQs_XS | RSs_M | DOW2 |
| ... | ... | ... | ... | ... | ... | ... |

**Table 3.2.** Sample of Tokenized Event Data

## 3.2 Analysis

Each network event was treated as a 'document' and each element of the network event, e.g. SKT80 was treated as a 'word'. The resulting corpus of 151,414 documents was analyzed using a simple LDA analysis. Collapsed Gibbs sampling was used under the assumption of $T = 50$ clusters, $\alpha = 0.01$, and $\beta = 0.01$.

The number of clusters as well as the initial values for $\alpha$, and $\beta$ were explored for sensitivity. Cluster sizes from 50 to 100 were investigated with no significant impact (other than computer cycle investment). Values of $\alpha, \beta = 0.01 - 0.1$ were explored. It was observed that as $\alpha$ became smaller the anomalous events were easier to identify, but the computational burden became cumbersome. However, regardless of the specific value for $\alpha$, the anomalous events were identified, either as individuals (low $\alpha$) or within a relatively small subset, e.g. 600 events, of the original 115,414 network events.

Figure 3.2 depicts plots of the 50 clusters that results from the analysis. Each plot represents a row of the $\Theta$ matrix. The x-axis of each plot is indexed by each of the 151,414 network events. The y-axis depicts the probability that the network event belongs to that particular cluster. For examle, if we pick a single x-axis value (a single network event) and sum the associated y-axis values across the 50 clusters the result is 1.0.

Counting left to right from the top left as Cluster 1 to the bottom right as Cluster 50, we see that a number of unusual patterns in network traffic appear. The vertical 'stripes' of, for example, Cluster 8, and the horizontal 'stripes' of Cluster 20, are atypical.

As noted in Section 2.2, to identify the anomalous network events various similarity measures were explored. The measure that provided the highest resolution of events was the Hellinger Distance. Recall that the distance measures involve a pair-wise comparison with the 115,414 documents; therefore the resulting Hellinger Distance matrix is a 115414 x 115414 matrix. For presentation purposes the Hellinger matrix is randomly down-sampled and only the down-sample results are presented in the following discussion. It should be noted that the sampling was biased to include documents (i.e. network events) that exhibited high Hellinger distance measures.

Figure 3.3 depicts the heat map that results from a Helllinger distance matrix of a random sample of 1000 random network events. The dendrograms on the edges highlight the clustering of network events into network usage patterns. The color scheme for the heat map was chosen to highlight extreme values in the Hellinger matrix.

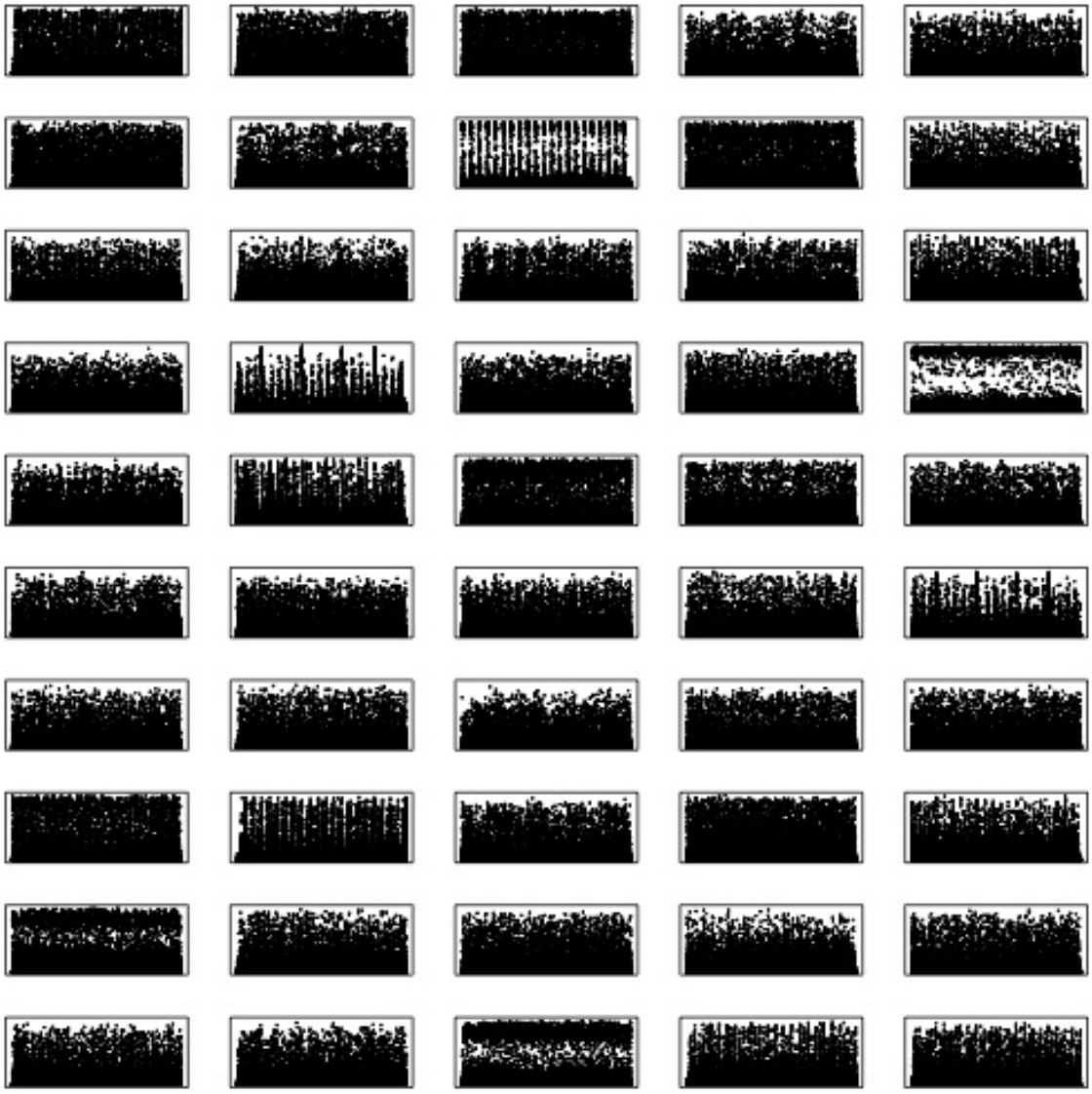

**Figure 3.2.** Resulting 50 Clusters

Figure 3.4 provides a better perspective on the results. In this case, the Hellinger distance matrix has been down sampled to a 500x500 matrix. It is now possible to identify nineteen anomalous network events (far left). The associated heat map is presented in Figure 3.5.

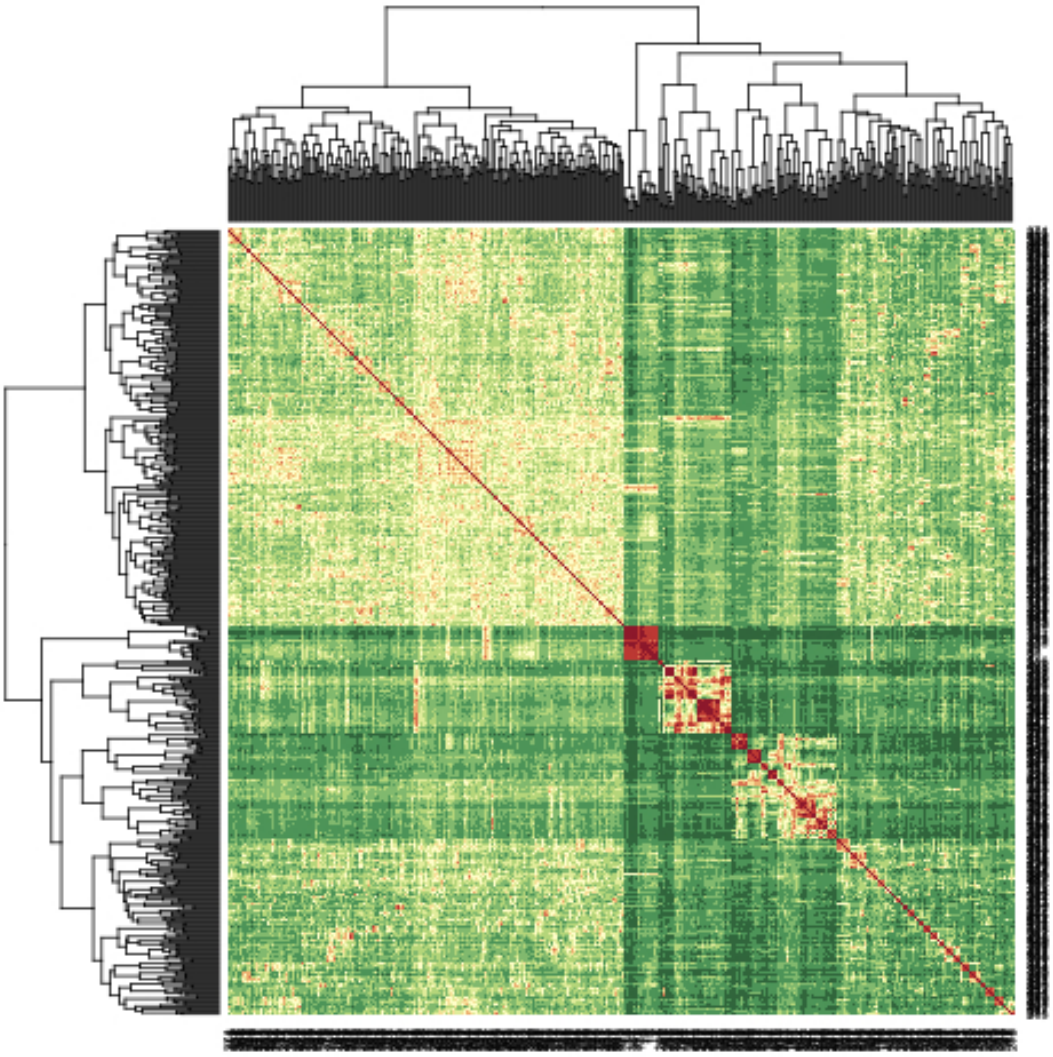

**Figure 3.3.** Heat Map of Hellinger Matrix - 1000 samples

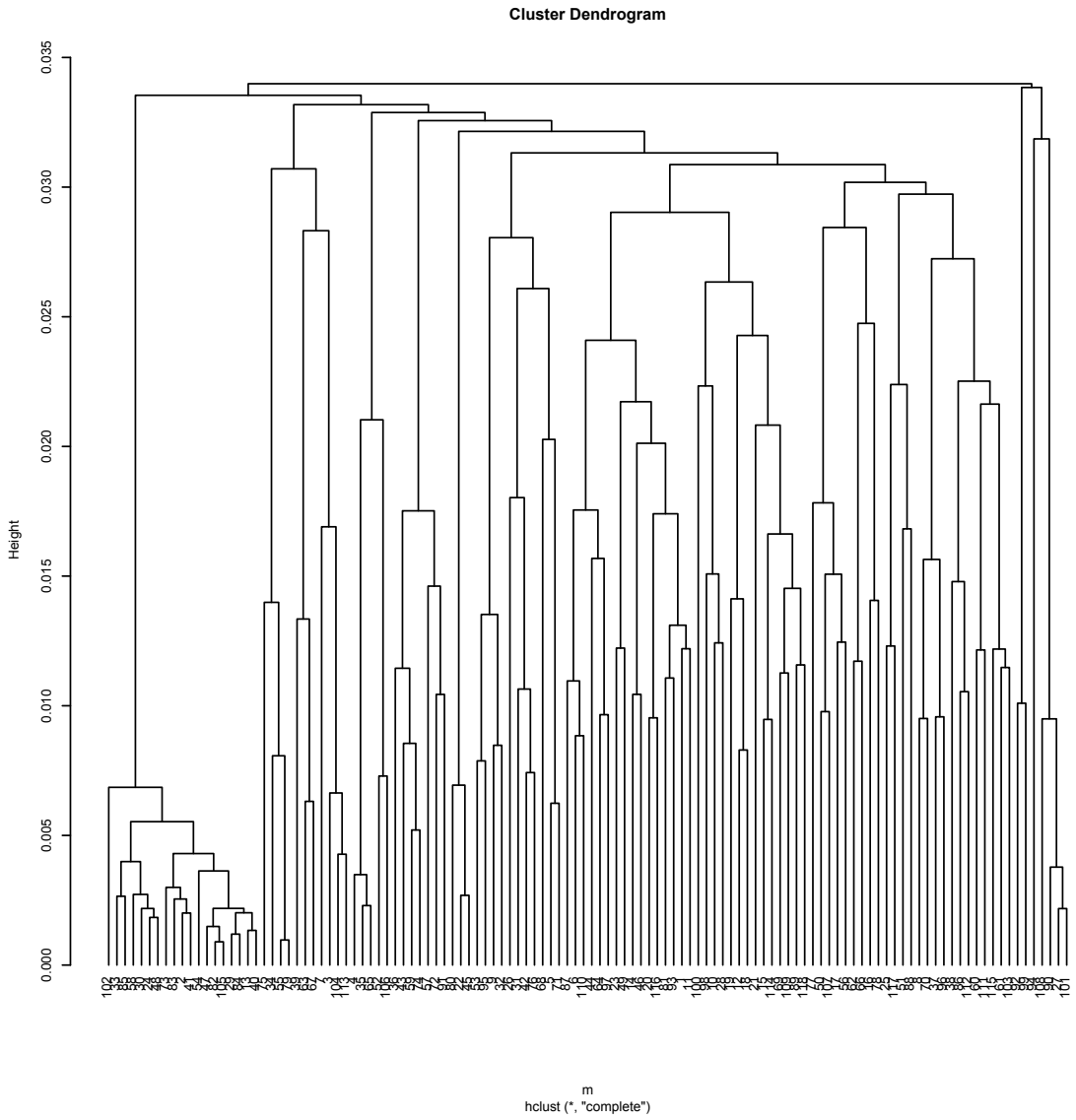


**Figure 3.4.** Hellinger Matrix Dendrogram - 500 samples

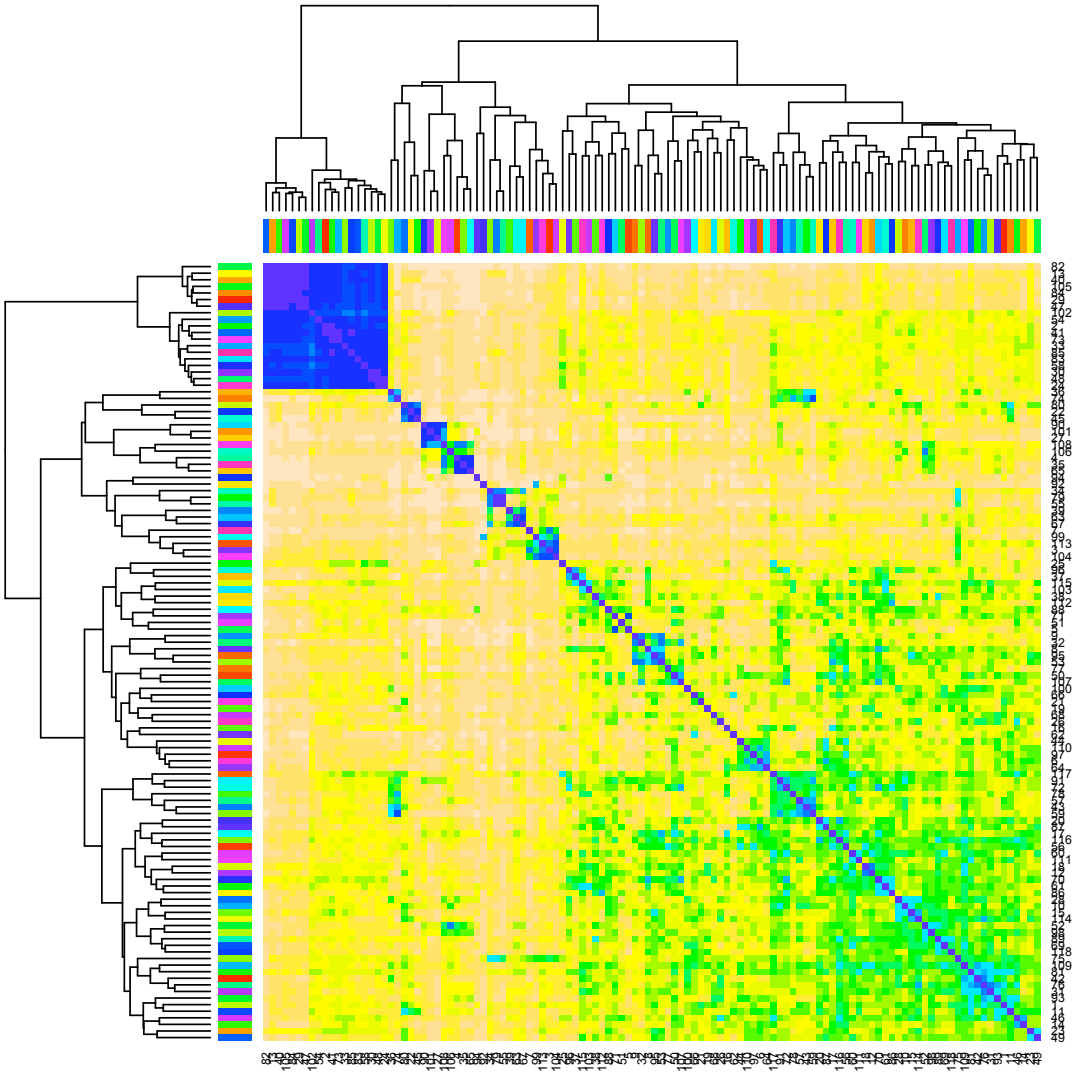

**Figure 3.5.** Heat Map of Hellinger Distance - 500 samples

# Chapter 4

# Results

Table 4 lists the nineteen network events associated with the red cells in the middle of Figure 3.3 and the branches on the far left of the dendrogram in Figure 3.4. (Similarly, the blue cells in Figure 3.5.) The nineteenth entry in the table is highlighted.

| | | | | | | |
|---|---|---|---|---|---|---|
| SIP3717010031 | TM17 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_XS | DOW2 |
| SIP3717010031 | TM14 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_XS | DOW4 |
| SIP3717010016 | TM16 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_XS | DOW4 |
| SIP3717010016 | TM16 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_XS | DOW2 |
| SIP3717010031 | TM17 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_XS | DOW2 |
| SIP3717010041 | TM12 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_XS | DOW4 |
| SIP3717010018 | TM17 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_S | DOW4 |
| SIP3717010013 | TM08 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_S | DOW2 |
| SIP3717010016 | TM17 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_XS | DOW2 |
| SIP3717010010 | TM09 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_XS | DOW4 |
| SIP3717010032 | TM10 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_XS | DOW4 |
| SIP3717010020 | TM17 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_S | DOW4 |
| SIP3717010056 | TM15 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_S | DOW2 |
| SIP3717010041 | TM16 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_S | DOW2 |
| SIP3717010020 | TM16 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_S | DOW2 |
| SIP3717010052 | TM09 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_XS | DOW4 |
| SIP3717010015 | TM13 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_XS | DOW4 |
| SIP371701008 | TM16 | DIP10059151133 | SKT8080 | RQs_XXL | RSs_M | DOW4 |
| SIP3717010031 | TM17 | DIP1031082140 | SKT8080 | RQs_XXL | RSs_XXS | DOW4 |

**Table 4.1.** Final Set of Suspect Network Events

Since the *truth* set was available from the VAST Challenge, the actual exfiltration events were used for validation. (The truth data set was not available at the beginning of the analysis.) The first eighteen network events identified by the linguistic analysis, and summarized in Table 4.1, correspond to the exfiltration events provided by the VAST Challenge.

The last entry in the table has some characteristics which sets it apart from other network events (e.g. use of socket 8080 and a very large requested packet size). However, the event also has some inconsistent characteristics, notably the destination IP address and the very small response packet size.

It should be noted that, while all the exfiltration events were all associated with the destination IP address of 100.59.151.133, these were not the only network events associated with that particular IP. In addition, the 8080 socket was used in other transactions.

Values of $\alpha, \beta = 0.01 - 0.1$ were explored. It was observed that as $\alpha$ became smaller the anomalous events were easier to identify, but the computational burden became cumbersome. However, regardless of the specific value for $\alpha$, the anomalous events were identified, either as individuals (low $\alpha$) or within a relatively small subset, e.g. 600 events, of the original 115,414 network events. The suggestion is that values of $\alpha$ be chosen to fit the needs of the organization. For example, to quickly reduce the number of suspect events down to a workable leve by system security staff, choose a larger value of $\alpha \approx 0.1$ and manually post-process the resulting list of suspect events. If a more specific identification of the events is desired, then use a smaller value of $\alpha \approx 0.01$ and process the network data off-line.

## 4.1 Discussion

Cyber security experts currently depend heavily on a rule-based framework for initial detection of suspect network events. The application of the rule set typically results in an extensive list of suspect network events that are then further explored manually for suspicious activity. The ability to identify anomalous network events is heavily dependent on the experience of the security personnel wading through the network log. Limitations of this approach are clear: rule-based systems only apply to exfiltration behavior that has previously been observed, and experienced cyber security personnel are rare commodities.

A key point is that the suggested analysis method can be accomplished in conjunction with the traditional, rule-based approach. Since the new methodology is not a discrete rule-based approach, it is more difficult for an insider to 'hide in the weeds' and disguise the exfiltration events. In addition, the approach can be extended to allow for a formal risk-based approach to attribution. Resources for investigation of possible exfiltration can be allocated based on the probability that there was one or more events. The approach is flexible in that in-line processing can be mixed with an off-line investigation and it is easy to update to account for changes in user network behavior (e.g. seasonal periodic pattern). Finally, the methodology provides an approach that can be implemented in a continuous, dynamic or evolutionary fashion. This permits suspect network activity to be identified early.

# References


[1] Blei, D. M. and J. D. Lafferty, 2007: A correlated topic model of science. *The Annals of Applied Statistics*, **1**(1), 17–35.

[2] Blei, D. M., A. Y. Ng, and M. I. Jordan, 2003: Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.

[3] Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, 1990: Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, **41**(6), 391–407.

[4] Hofmann, T., 1999: Probabilistic latent semantic analysis. *Uncertainity in Artificial Intelligence, UAI'99 Stockholm*.

[5] Li, W. and A. McCallum, 2006: Pachinko allocation: Dag-structured mixture models of topic correlations. *ICML '06: Proceedings of the 23rd international conference on Machine learning*.

[6] Meila, M., 2003: Comparing clusterings by the variation of information. *16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003,* 1–15.

[7] Robinson, D. G., 2009: IEEE VAST Challenge: www.cs.umd.edu/ hcil/ vastchallenge09.

# DISTRIBUTION:

| | | |
|---|---|---|
| 1 | MS 0671 | Tan Chang Hu, 5627 |
| 1 | MS 0671 | Mitch McCrory, 5627 |
| 1 | MS 1316 | Scott Mitchell, 1415 |
| 5 | MS 1318 | David Robinson, 1415 |
| 1 | MS 9159 | Philip Kegelmeyer, 8962 |
| 1 | MS 0899 | Technical Library, 9536 (electronic copy) |

Sandia National Laboratories