UNCLASSIFIED

Australian Government
Department of Defence
Science and Technology

Bayesian Modelling of Network Traffic Metadata using Dirichlet Multinomial Mixtures

Kevin Harman

Cyber and Electronic Warfare Division
Defence Science and Technology Group

DST-Group-TR-3538

ABSTRACT

Statistical theory commends probabilistic modelling techniques for the discovery of latent structure in large datasets not amenable to analysis by inspection. Network traffic metadata, for example, may contain latent structure representing different traffic behaviours. The utility of a class of Bayesian models known as Dirichlet multinomial mixtures in discovering such behaviours, and how they might be applied to network analysis problems such as source characterisation, event detection or filtering, is considered herein. Encouragingly, under the right conditions, these models are found to detect and quantify meaningful behavioural distinctions. For an analyst tasked with understanding unpredictable and rapidly evolving traffic, but limited by privacy, volume or encryption to abstract data, behavioural learning like Dirichlet mixture modelling could prove a valuable tool.

RELEASE LIMITATION

Approved for public release

UNCLASSIFIED

Produced by

Cyber and Electronic Warfare Division
Defence Science and Technology Group
PO Box 1500
Edinburgh South Australia 5111 Australia
Telephone: 1300 333 362

© Commonwealth of Australia 2018
October 2018

APPROVED FOR PUBLIC RELEASE

(U) Bayesian Modelling of Network Traffic Metadata using Dirichlet Multinomial Mixtures

Executive Summary

In the field of network traffic analysis, constraints including privacy, encryption and capacity give impetus to a transition from analysis based on deep packet inspection to analysis based on *metadata*.

Whereas packet inspection provides full visibility of (unencrypted) content and therefore low ambiguity in interpretation, metadata is content-opaque, and even dissimilar transactions might have near identical metadata records. Communications once characterised by reference to known byte sequences, or *signatures*, must instead be assessed by trends, or *behaviours*, and the high ambiguity in metadata demands such trends be inferred over multiple observations. There arises an issue of scale, both from this volume and high dimensionality in the metadata, so that the data is no longer amenable to analysis by inspection.

Instead, scientific theory commends *statistical learning* for the discovery of latent structure in data. Bayesian probabilistic modelling techniques from this class, although well established in many other domains, are only recently emerging in network analysis. A subclass known as *Dirichlet multinomial mixture* (DMM) models appears particularly well matched to network problems, describing a structure in which multiple disparate sources of data are mixed together at measurement, much as the modern internet mixes many disparate protocols and services on a common transport infrastructure. Accordingly, this report seeks to assess the utility of DMM modelling with network metadata in roles such as source characterisation, detection of cyber security events, or related filtering. The significant output from the model is a description of each identifiable source, providing two derivative results - a clustering of data by source, and a measure of likelihood that data should belong to a source.

From a broad range of potential research activities identified, this work concentrates on assessing DMM against filtered views of highly aggregated internet backbone traffic and with a variety of data attributes. The major outcomes are:

- DMM is a suitable model choice for network traffic metadata, i.e. the model building process should converge, producing a manageable number of distinct sources, each of which can typically be explained by behavioural trends.
- There is the potential to use a broad range of attribute combinations to describe network data observations, and this choice can significantly alter the modelling outcomes. Attributes may be literal metadata fields or derivations thereof. Correlation between attribute combinations should be minimised to avoid lack of resolution in the source descriptions.

- Model building was effective against the traffic in both highly aggregated and tightly filtered forms.
- Trends in data clusters per source can assist characterisation and detection.
- Trends in likelihood measurements can also be related to behaviours.

Statistical learning has relevance to the Australian Signals Directorate (ASD), which is the entity charged with provision of Australia's Signals Intelligence (SIGINT) capabilities. ASD analysts must use metadata to mitigate privacy, volume and encryption constraints. To deal with the information loss of metadata abstraction and the unpredictable behaviours of cyber adversaries with ever-evolving tools, techniques and procedures, unsupervised learning like DMM modelling must form the foundation of future toolsets.

Author

Kevin Harman
Cyber and Electronic Warfare Division

Kevin was awarded a Bachelor's Degree with Honours in Electrical & Electronic Engineering from the University of Adelaide in 1991. He joined the then DSTO in 1998 and conducts research and development in communications systems.

This page is intentionally blank.

Contents

Glossary

ASD	Australian Signals Directorate
C2	Command and Control
CNR	Communications Network Research (Group)
Cnt	Count
CSD	ContinuStor Monitor
DMM	Dirichlet multinomial mixture (model)
DNS	Domain-name Server (or service)
Dst	Destination
DstIP	The address of the destination host in IP-based network traffic
DST Group	Defence Science and Technology Group
DstPt	The port expected at the destination host in internet protocol-based network traffic
EM	Expectation-maximisation (solving technique)
HTTP(S)	Hypertext Transfer Protocol (Secure)
ICMP	Internet Control Message Protocol
IP	Internet protocol, for management of network traffic, or specifically the address of a host in such a network
IPFIX	IP Flow Information Export
Loglik	Log-likelihood
MCMC	Markov-chain Monte-Carlo
MQtt	MQ Telemetry Transport (protocol)
MS	Microsoft Corporation
NTP	Network Time Protocol
Oct	Octet (or byte)
Pkt	Packet (of network traffic)
POA	Point-of-access (for network data measurement)
RDP	Remote Desktop Protocol
SIGINT	Signals Intelligence
SIP	Session Initiation Protocol
SrcIP	The address of the source host in IP-based network traffic
SrcPt	The port nominated by the source host in internet protocol- based network traffic
SSH	Secure Shell protocol
SSL	Secure Socket Layer
TCP	Transmission Control Protocol, a connection-oriented transport-layer protocol
TTL	Time-to-live
XMPP	Extensible Messaging and Presence Protocol

This page is intentionally blank.

1. Introduction

1.1. Motivation

Network traffic is routinely monitored to maintain performance, reliability and security [1]. The highest visibility of content, and hence the lowest ambiguity in analysis, is afforded by unencrypted, raw packet data, for which both measurement and security tools are abundant [2]. However constraints including privacy, encryption and capacity often limit the utility or accessibility of packet data. In these situations tools and analysts often rely instead on *metadata,* a summarised and abstracted record of packet-level communications, and in consequence accept more ambiguous conclusions.

With metadata, characterisation and cyber security tasks that were solved by referencing literal signatures become instead tasks of searching for structure or patterns in the data and attributing1 these to network aspects of interest. Unfortunately, large, abstract datasets are rarely amenable to analysis by inspection.

Instead, scientific theory commends *statistical learning* for the discovery of latent structure in data [3]. Bayesian probabilistic modelling techniques from this class [4], although well established in many other domains of analysis, are only recently emerging in network analysis tasks [e.g. 5]. Further, the particular class of models known as *Dirichlet multinomial mixture* (DMM) models [6] appear well matched to network problems, describing a structure in which multiple disparate sources of data are mixed together at measurement, much as the modern internet mixes many disparate protocols and services on a common transport infrastructure.

Accordingly, this work seeks to evaluate the utility of DMM modelling of network metadata in roles such as source characterisation, detection of cyber security events, or volume filtering in support of same.

1.2. Traffic Metadata Concepts

Metadata representations of Internet protocol (IP) traffic [7] summarise sets of related packet exchanges between two network end-points as a single *flow* or *session* record. Industry-standard formats include Netflow [8] and IPFIX [9]. Herein the terms flow, session or Netflow are used interchangeably with the same meaning.

The core attributes of a session relate to connectivity (hosts and ports), payload size (packet- and byte-counts), time (date-stamps for the beginning and end of each session), and protocols (TCP/IP flags, etc.). Derived attributes are also possible. The Netflow record used in this research has the form shown in Figure 1.

[1] Note that according to these *tipoffs*, targeted packet capture for more precise resolution might be warranted.

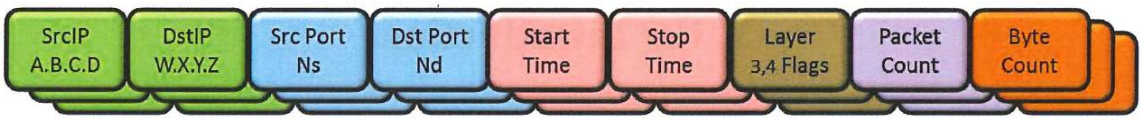

Figure 1: Data fields in simple Netflow

Here:

- SrcIP is the IP address of the host which initiated the session (nominally a *client*[2]), and SrcPt is the port opened for returned packets. *IP version 4* addressing is assumed throughout and has a 4 gigabyte range. Port allocations are in the discrete range 0 to 65535.
- DstIP is the IP address of the host to which the communication is directed (nominally a *server)* and DstPt is the allocated port for the provided service.
- Packet and Byte counts reflect the total numbers of packets and bytes exchanged during the session.
- Start and Stop times are a full date-stamp of the beginning and end of the communication session between the two hosts, typically with millisecond precision.
- Other fields are ignored.

1.3. Statistical Learning Concepts

Statistical learning is the application of mathematics from the domains of probability and statistics to generating information or knowledge from data [3]. Such knowledge might include a *model* which describes attributes of the data or of the sources perceived to have generated it, usually in the form of parameters (variables) for which the model provides numeric estimates. The model will often allow the intuitive notion of *likelihood* to be expressed in a probability equation. Then, natural language queries such as 'is this an unusual member of the dataset' can be evaluated numerically. These techniques are most beneficial in so-called *big data* problems, where scale issues render improbable the detection of useful patterns or knowledge by inspection alone.

Netflow from backbone traffic qualifies as 'big data'. Significantly, Netflow also qualifies as *sparse* data. In sparse data, the elements in each record can take wide-ranging values, with the result that even in a large sample of data it is extremely unlikely all possible values the data could take (in all combinations) will occur. Wide parameters such as internet-protocol (IP) address, and source or destination port (SrcPt, DstPt), as well as limitations in network visibility at a given point of access (POA), ensure Netflow is sparse.

[2] As a function of the Netflow collector and access posture, some sessions may not be detected until after the initial packet, allowing the destination side to be detected as an originator. These events are simplistically managed in this work by mapping the server-side in accordance with the smaller port value.

This is problematic in some forms of analysis in that new observations that were not present at the time of model building are declared to have zero-probability of occurrence, with implications for false-alarm rates in detector applications.

Probabilistic modelling techniques are particularly well suited to modelling Netflow:

- They produce models which profile the data sources at the POA, facilitating both change detection and outlier detection even in a big-data context.
- They produce smooth models which are robust to the occurrence of data samples that were not present when the model was built. In this way, they are resilient to sparseness.
- They are from the class of machine learning tools known as unsupervised, meaning the models can be built from the data alone, requiring no prior knowledge of what constitutes normal or outlier traffic.
- Nonetheless, if prior knowledge about the data sources exists, it may be incorporated in the model effectively.

1.4. Dirichlet Multinomial Mixture Model Concepts

Under the assumption that data to be modelled was generated from a finite number of distinct *sources* that have been *mixed* together, DMM modelling [6] seeks to segregate the data back into its original source clusters and provide a description of each source. These descriptions are in the form of estimates of probability distributions, samples from which would mimic real data samples from the same source.

The source clusters may be used to *profile* the data composition, and the source distributions establish a mathematical framework for predicting what new observations from the same sources might look like, or how well any given observation fits the model.

With flow data from telecommunications networks, source profiling can provide insights into the expected composition of traffic at a POA (e.g. proportions of HTTP, DNS, Mail, P2P, scanning, etc.), and the predictive framework facilitates anomaly detection by providing differentiation between likely and unlikely traffic.

A DMM model may be depicted as a Bayes net, as in Figure 2 [6]. The Bayes net identifies both the variables in the model (the nodes) and the manner in which they are conditionally dependent on each other (the directed arrows).

The variables in this model include:

- $\mathbf{X}$ *(the data or observation)*. In this work, each observation is a *feature vector* derived from a set of related Netflow sessions, and the number of dimensions p in each observation is the number of fields in the feature vector. There are usually many (n) such observations.

- **θ** *(the source distributions)*. Observations *X* are assumed to have been generated from one of several sources (of which the actual number *S* is hypothesised during model building). A vector θ_s exists for each source as a probability distribution for each dimension of *X*.

- **ϕ** *(the source proportions)*. The *S* sources are assumed to occur in different proportion in the dataset. ϕ defines the fractional proportion of each.

- **Z** *(the source selector variable)*. *Z* is a marker which indicates which source, in the range 1-to-S, the corresponding sample of X was generated from. If there are *n* observations, there are also *n* values in *Z*.

- α *(the prior estimate of θ)*. α has the same dimensionality as θ. This so-called prior can be used to bias the way θ is estimated during the model building stage if some aspects of the source distributions are known or expected in advance.

- β *(the prior estimate of ϕ)*. ,β has the same dimensionality as ϕ. This so-called prior can be used to bias the way ϕ is estimated during the model building stage if some aspects of the source proportions are known or expected in advance.

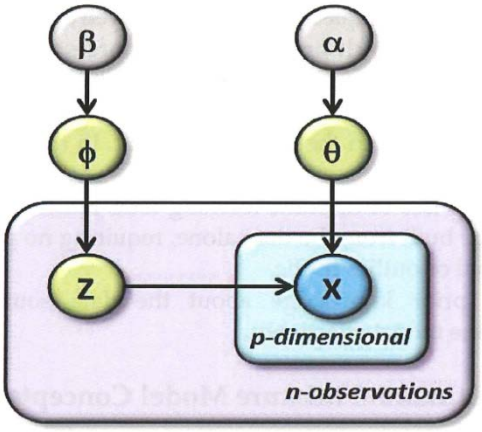

Figure 2: The Bayes Net description of a DMM model

The gist of the model is then intuitive:

> In order to understand what a sample (***X***) from the dataset should look like, it is necessary to know both which source the sample is from (***Z***) and what data from that source looks like (***θ***).

The choice of source itself depends on the relative proportions of the sources (ϕ). Finally, the parameters α and β are known as priors. Where the data lacks evidence to adequately describe ϕ or θ (for example due to sparseness), but the modeller expects certain behaviours, the priors provide a mechanism to bias the source descriptors toward the expected outcomes.

For completeness, from [6]:

- The general formulation of a Bayesian problem asserts that the chance that the model (or *hypothesis* about the set of parameters and their values) takes a certain form, given the data that has already been observed, is proportional to both the likelihood of seeing that data, given the model hypothesis, and the likelihood of that set of parameters occurring, i.e.:

$$P(Hypothesis|Data) = \frac{P(Data|Hypothesis) * P(Hypothesis)}{P(Data)}$$

  Where
  - the *hypothesis* is the set of parameters that describes the model
  - P(Hypothesis | Data) is known as the *posterior* distribution
  - P(Data | Hypothesis) is known as the *likelihood*
  - P(Hypothesis) is known as the *prior*
  - P(Data) is known as the *evidence.*

- This is equivalently expressed as:

$$P(\Theta|D) = \frac{P(D|\Theta * P(\Theta)}{P(D)}$$

  Where
  - Θ is the vector of model parameters
  - D is the set of observations.

- In the specific case of DMM, the expression for the Posterior probability becomes:

$$\mathrm{P}(\emptyset, \theta, \mathrm{Z}, \mathrm{X}\,|\alpha, \beta) = \frac{1}{C(\beta)} * \prod_{k=1}^{l}\left(\emptyset k^{\beta k-1+Nk(Z)} * \frac{1}{C(\alpha)} * \prod_{j=1}^{m}\left(\theta kj^{\alpha j-1+\sum_{i:Zi=k} Nj(Xi)}\right)\right)$$

  Where:
  - $C(\alpha) = \frac{\prod_{j=1}^{m} \Gamma(\alpha j)}{\Gamma(\sum_{j=1}^{m} \alpha j)}$
  - Γ is the Gamma function
  - *N* are *multinominal* observations of subsets of X.

1.5. Document Scope

Applying DMM modelling to Netflow and assessing the outcomes is a broad task. Section 2 imagines the full scope of this research, and identifies the subset of work undertaken herein.

Test data was Netflow collected at the level of a tier-2 service provider [10]. Section 3 examines the composition of this data, applies a filtering scheme to extract a controlled subset for testing, and identifies deliberately *seeded* data added to allow exploration of cyber-security aspects.

Then in Section 4, a DMM model is built from the filtered data and examined. Basic properties of model convergence, source composition and likelihood interpretation are addressed with respect to the generic goals of source characterisation, cyber event detection, and filtering.

Section 5 seeks to extend via the important aspect of *feature selection*, which maps the raw data to samples, or *observations*, in the model, with implication for the range of discernible *behaviours*. A potential weakness of ad-hoc feature selection is exposed, but redressed in Section 6.

Section 7 stands alone in consideration of model building for select subsets, or *domains*, of the traffic.

Finally, Sections 8 and 9 consider how modelling might proceed from research toward an analyst capability, and offer concluding remarks.

2. Research Scope and Coverage

The task of assessing the utility of DMM modelling against Netflow can be split into the three broad categories of *data management, model building* and *model analysis*.

Within data management are the subtasks of *collecting* the data, *labelling* the data with useful attributes from content inspection or other analysis tools, *seeding* the data with reference traffic that may aid in later analysis, and possibly *filtering* the data to a specific class or domain in order to build a targeted model.

Model building includes stages of observation construction, which addresses *sampling* and *feature selection*, *implementation* as executable code including validation, the specification of *priors*, model *solving* and model *selection*.

In the analysis stage model outputs, which for DMM include source descriptions, cluster assignments and a *likelihood* equation, are assessed for utility in characterising the data source, detecting outliers or filtering traffic classes of interest.

Collectively these tasks create a substantial test surface, comprehensive coverage of which is beyond the scope of this report. Instead, Table 1 below provides additional commentary on the scope of each aspect and indicates to what extent that scope is tested in this research.

Table 1: DMM Modelling Research Concepts and Coverage.

Task		Discussion
Collection	Scope	At different points-of-access (POA) to a network, a different extent of traffic and devices may be visible; varying what may be inferred from analysis. Factors such as access to packet-level data, collection constraints due to bandwidth or time-of-day, the type and diversity of traffic carried by the network, the collection sensor location, etc., will constrain feature augmentation, model development and analysis. Multiple features should be made available, including from the categories of *identity, connectivity, size, time* and *protocol.* Multiple access points should be considered including from the types backbone, enterprise gateway and stub network.
	Coverage	This research uses basic Netflow collected at the level of a tier-2 internet provider, with a highly diverse range of addresses and traffic types available. Details are provided in Section 3. A subsample of *heavy-hitters* is extracted for Sections 4, 5 and 6, and there is limited application to domain-specific subsets in Section 7
Labelling	Scope	Labelling associates a useful descriptor with each data record. For supervised learning schemes, accurate labelling is essential. With unsupervised learning, interpretation of results is subjective without relevant labels. Labelling could address: • Identification of client and server nodes in a session. • Classification, e.g. by protocol or application. • Sessions (e.g. an application tag) or nodes (e.g. web-server, mail-server). The latter labelling may also be referred to as *device characterisation,* which must also manage multi-f unction nodes.
	Coverage	Client/ server demarcations were approximated by association of the small session port with the server-side. Per-session application labelling by port-based classification was used [11]. Reliability is expected to be on the high side of such approaches based on the pre-filtering described below. However, due to the behavioural nature of the modelling, a direct correlation between clusters and service ports was not assumed.
Seeding	Scope	Labelled traffic displaying known behaviours may be inserted into the data to influence model building and analysis. This is particularly useful in a detection context, where demarcation of the seeds can be tested, or in a characterisation or filtering context where seeds impart their behavioural attribute to co-clustered traffic. The types and proportions of seed traffic that can be added without *biasing* model building are unknown.
	Coverage	Two classes of seeded traffic are inserted into the original data. One is emulated traffic from a host infected by malware that is communicating with a remote command and control (C2) server. The second is Netflow alerted by an intrusion detection

		system (IDS). An attempt is made to limit bias by maintaining a high ratio of non-seed to seed traffic. See Section 3.
Filtering	Scope	Filtering the data before use could take many forms, including: • Sampling (regular or random), e.g. to manage volume. • Selection by subnet(s) (including geographic zones). • Selection by service(s). • Selection by label, including labels from other statistical learning operations. • Whitelisting. There are implications for (at least) completeness (especially where aggregation is required to explain a behaviour), efficiency, bias and accuracy.
	Coverage	The original data was filtered to a selection of heavy-hitters, preserving measures of server port composition and host diversity. In some tests, additional filtering to either seed-only or service--only traffic was also applied.
Code Development and Validation	Scope	There are multiple choices for the implementation of the modelling algorithm, with implications for development cost, efficiency, scalability, portability, compatibility, adaptability, etc. The use of second-party, open-source code reduces the validation burden and development time, but may limit control and monitoring during fundamental investigation of the algorithms. The use of custom code presents the converse scenario.
	Coverage	The DMM algorithm, as transcribed from [6], and associated data processing and analysis were implemented in Matlab script [12], facilitating rapid prototyping and stepwise querying of the modelling. Pseudo-code is provided in Appendix A. Validation of model convergence and source separation was tested using mixtures of Normally distributed sources.
Priors	Scope	Prior probabilities allow the model to handle sparseness robustly, and influence model convergence from prior knowledge where evidence in the data is lacking or a specific bias in the model is sought.
	Coverage	Due to both a lack of insightful prior knowledge and the intention to ensure model outputs reflect only the behaviours of the raw data at hand, uniform priors were used.
Feature Selection	Scope	Feature selection is the critical task of choosing the subset of data attributes that the model will use to segregate behaviours. Basic Netflow offers features in the classes of *identity*, *connectivity*, *timing*, *size* and *protocol*. Additional features may be added by derivation that might relate to (at least): • averages or higher moments; • rate of change; • packet attributes such as size or time distributions; • application or class labels, e.g. from prior statistical learning;

		• self-similarity; and • functions or transformations (e.g. Fourier transforms). Ideally, the *combination of dimensions* should be chosen consistent with the intent of the modelling, although the impact of selection on modelling outcomes and the domain knowledge required to inform such selection may not always be known. These heuristic choices may require systematic moderation, for example to eliminate dimensions that are highly correlated and hence expand the model without enriching the information content. Note that with DMM modelling, since observations take the form of multinomials (i.e. counts across a finite set of possible outcomes), there is no need for parameterisation of distributions - the *empirical histogram* is the required observation, and all information in the distribution is passed to the model provided the class intervals have sufficient resolution.
	Coverage	Initial modelling is literal, seeking separation of behaviours by providing a direct representation for ·each of the fundamental feature classes in Netflow. These absolute-value representations are then supplemented with difference and self-similarity measures to seek enrichment, a solution which is found to require moderation by correlation testing. Feature selection is addressed in more detail in Sections 4 to 6
Observation Construction	Scope	The measured data elements, in this case Netflow records, must be mapped into *observations* in the model. This may be a one-to-one mapping, or may be subject to rules of aggregation. For example, to reach a *behavioural* conclusion about a host or node in the network, a single session from that host is likely to be insufficient - due to the summarising nature of Netflow many different hosts and many different services will at times be indistinguishable from the perspective of a single Netflow record. Aggregation strategies could be referenced to an edge, a node, a service or class, etc. A minimum count may be required to ensure representative statistics. Aggregating by node will result in a smaller initial dataset which may simplify computational aspects at the expense of loss of behavioural resolution due to averaging. Recursive profiling could be used, e.g. for edge-based profiling of a node-based cluster. Aggregation could be windowed by block size or time, achieving recurring observations of the same entity. The set then allows patterns related to sequence or multiple behaviours to be included, with respect to the resolution of the chosen window. In representing an aggregated set of values that does not conform to a standard parameterised distribution, a histogram is often required, for which a set of class intervals must be chosen. The resolution of these bins must be chosen with

		respect to the behaviours to be elucidated - low resolution creates an averaging that may mask some trends, and high resolution increases the computational burden during model evaluation
	Coverage	In a majority of tests observations are per-host aggregates for the server-side of the transaction. Hence the models attempt to characterise the behaviours of the set of servers in the data. A minimum of twenty sessions are required per observation, based on prior empirical evidence of histogram stability using data from the chosen POA. Select tests use a block-based window per-host, with block size of twenty. Dimensions are binned over linear ranges that encompass their empirically observed minimum and maximum values, although dimensions with wide ranges may first be logarithmically (base-2) compressed.
Model Solving	Scope	Model solving is the task of fitting the model to the observed data and can take a variety of computational forms such as expectation-maximisation (EM), Markov-chain Monte-Carlo (MCMC) sampling, variational inference, etc. [13]. In the DMM model, Dirichlet-conjugacy allows the model equations to be expressed in a form most suitable for the Gibbs-variant of MCMC solving [6]. Solving must also bound or control model dimensionality. These solvers are iterative and hence require decisions on convergence and stopping points.
	Coverage	Gibbs MCMC solving was applied. The source count dimension S is unknown a-priori. The model building approach adopted was to assume a large value for S and accept this if the result includes null-sources. Source proportions ϕ were used to indicate convergence, with stability in proportion estimates typically achieved within 200 samples. A further 800 samples were taken as estimates from model variable posteriors.
Model Selection	Scope	Given the output of solving as a set of distribution estimates for random variables, model selection is the task of sampling the joint space for the best representation with regards to a particular goal or requirement, e.g. to find point estimates of the model space from which to generate new samples of data. This is an expansive problem [13]. Depending on the class of Bayesian formulation, this may involve selecting structure, variables, or both.
	Coverage	In the DMM case, model structure is assumed and only model variables require selection. Point-estimates of parameters were taken as the maximum a-posteriori (MAP) values of their estimated distributions.
Source Distribution Analysis	Scope	DMM source variables define the number and nature of distinct sources evident in the data. Sources could be investigated directly by the structure of their distributions, e.g. to test whether separation by Bayesian estimation relates to separation by Euclidean distance.

		Sources could be sampled to generate new observations.
	Coverage	Source distribution analysis and generative testing were not addressed.
Cluster Analysis	Scope	The indicator variable Z allows observations to be clustered by their originating source. This could be used to correlate how each source distribution, or *behaviour*, relates to attributes or labels of co-clustered observations.
	Coverage	Clusters are described subjectively by their majority labels and trends in the Netflow dimensions for connectivity, size and timing. This allows consideration of merit for characterisation, and using seed observations, for filtering and detection.
Likelihood Analysis	Scope	A numeric likelihood may be assigned to each observation with respect to the selected model. Depending on the extent to which the model probability equations can be solved analytically, this may be an absolute or relative assignment - for example when the *evidence P(X)* is not known, proportionality applies for the posterior probability, and the likelihoods can only be known up to a normalising constant [6]. Observations can then be ranked by likelihood, leading to identification of the least- and most-likely members, with potential to inform detection and filtering, e.g. were the seed observations distinct with respect to a likelihood threshold? The overall distribution of likelihood values may be examined, with the potential for modality to inform filtering or detection. Finally where there are multiple observations related to the same entity (e.g. a particular server IP), there exists a set of likelihoods which in aggregate pattern or in sequence may constitute a signature of interest for detection or filtering. Such vectors of likelihoods provide a mechanism for attributing multiple behaviours to the same entity. These might be tested by distance (clustering or similarity) or by distribution (e.g. a tendency toward certain quartiles of occupancy, or similar).
	Coverage	Least- and most-likely observations are identified. The distribution of seeds with respect to non-seeds is considered. Modes of the likelihood distribution are checked for commonality of behaviour. Patterns of likelihood-per-host are visualised

3. Data Management

3.1. Raw Data

Netflow for this research is in the category of highly diverse, aggregated backbone traffic, collected at the tier-2 level of internet infrastructure [10]. Some 16 billion sessions were available, each having the dimensionality shown in Figure 1.

Diversity is illustrated in Figure 3, showing the distribution of the SrcPt across the entire dataset (the distribution for DstPt is similar).

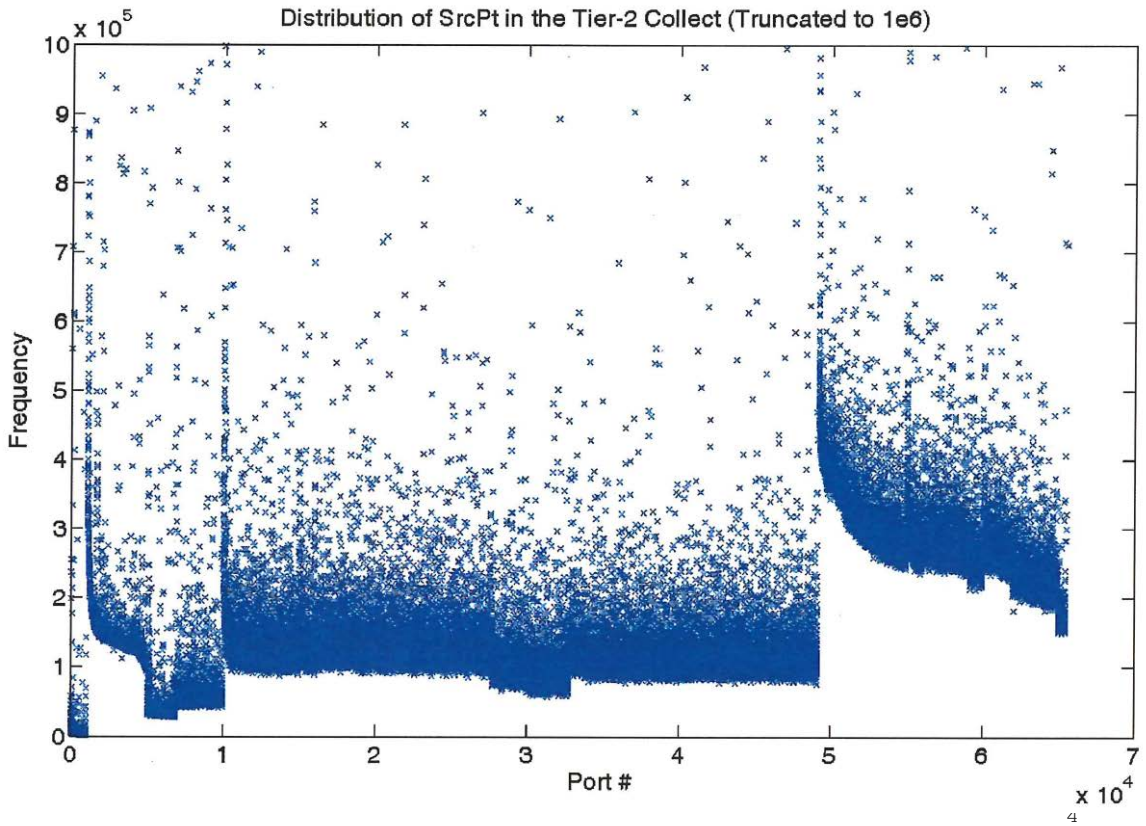

Figure 3: SrcPt distribution (truncated) of tier-2 backbone data

In this data, some sessions may not be detected until after the initial packet, allowing the destination side to be detected as an originator. Such events are simplistically managed in this work by mapping the server-side in accordance with the smaller port value.

3.2. Filtering

DMM model development was conducted on desktop-grade workstation hardware. This necessitated a reduction in the number of data records, as the full set has comparable count to the memory capacity of a workstation, and (based on empirical evidence) the

highly iterative nature of the computations results in slow execution time with standard serialised processing.

A *filtering* approach to set reduction (rather than a sampling approach) was adopted herein. Filtering has the potential to bias the dataset in subtle ways, which could lead to the modelling of behaviours that are not actually present. In order to minimise such effects, the approach adopted here was to reduce the session count while preserving both the relative distribution of *dominant* service ports and the relative distribution of *unique hosts* per retained service port (for a given sample size, number of ports and number of hosts per port, and counting from the most frequent to less frequent). It is acknowledged that by retaining these dominant ports and frequent *heavy-hitters* the dataset will be biased toward popular servers of standard services and their client sets - this could make the behaviours presented to modelling more distinct than they might otherwise be in the presence of a myriad of less common transactions.

Somewhat arbitrarily, a sample was taken such that:

- Size was of the order of 10 million sessions, known to be manageable in the available computing infrastructure.
- The top 33 SrcPt and the top 33 DstPt were used in equal measure, and known to account for all significant non-ephemeral ports in use. These form a set of 47 different 'service' ports in total.
- The top 20 hosts per port were retained.

Revisiting the data with this filter yielded 9 058 815 sessions with retained port distribution as shown in Figure 4.

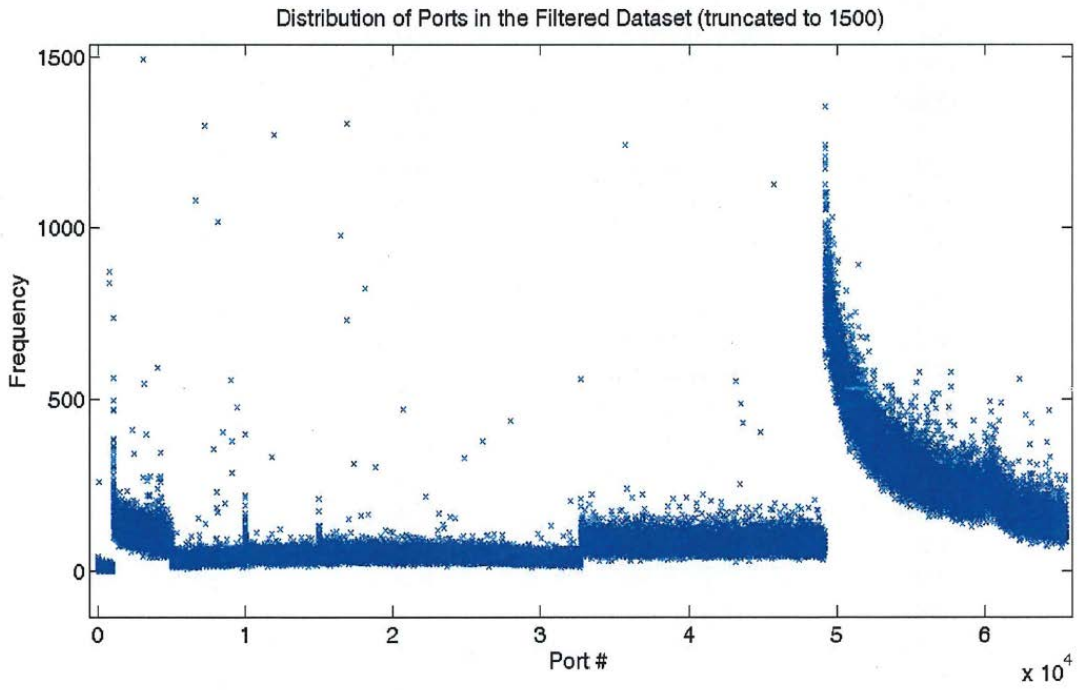

Figure 4 - Retained port distribution (truncated)

Table 2 provides a further breakdown of traffic composition. Here, the *Label* is by port classification, which is expected to be robust for the heavy-hitter subset. Labels in italics reflect ports with unofficial allocations or where the class was inferred from end-points in the filtered dataset.

Table 2: Class-based Composition of the Test Data

Nominal Service Port	Label	Session Count	Host IP Count
80	HTTP	2 752 186	18
443	HTTPS	2 431 080	11
53	DNS	976 389	22
0	ICMP	457 836	26
6881	Bittorrent	438 628	27
5223	XMPP	321 569	36
771	*Port Unreachable*	183 528	11
445	MS DS	161 213	5
3389	MS RDP	156 023	26
6000	Xll RDP	129 037	10

Nominal Service Port	Label	Session Count	Host IP Count
1433	MS SQL	97 924	20
25	SMTP	87 346	16
2048	*Echo Request*	85 685	20
22	SSH	82 623	14
4672	*eMule*	81 325	6
1024	*Bittorrent*	74 015	22
769	*Host Unreachable*	63 706	13
6890	*Bittorrent*	35 075	20
12350	*Skype*	32 607	16
8080	HTTP Proxy	31 008	19
23	Telnet	30 326	14
6882	*Bittorrent*	28 868	22
123	NTP	23 294	11
993	IMAP SSL	22 432	16
3306	MySQL	21 166	16
3072	*CSD Monitor*	20 996	19
781	*Policy Block*	20 382	12
4244	*Viber Desktop*	19 462	20
5060	SIP	15 919	9
2816	*TTL Expired*	15 351	12
15831	*Unknown*	13 412	4
135	MS RPC	12 980	8
995	POP3 SSL	12 878	6
8883	MQTT SSL	12 467	18
1025	MS RPC	12 441	8
110	POP3	12 200	16
4935	*FS RDP*	10 065	14
3074	XBOX Live	9411	7
12200	*X11*	8878	2
1026	*Unknown*	8823	9
5242	*Viber*	7442	20

Nominal Service Port	Label	Session Count	Host IP Count
8000	Radio	6907	10
5222	XMPP	6319	11
10000	Webmin	6065	13
5228	*Google Play*	5955	12
1027	*MySQL*	5952	6
6969	*Bittorrent*	5881	5

3.3. Seeding

The raw data was seeded to allow consideration of model efficacy in detection and filtering contexts. In these tests, seeded sessions were chosen to reflect the behaviour of malware infections [14, 15], which might be distinct to some extent by cluster or likelihood. Four different seed types were added as traffic between infected hosts and their presumed command and control (C2) next hop:

1. *SANS Seed:* This traffic is from the SANS Institute training course FOR572 (Advanced Network Forensics and Analysis) [16]. It describes an emulated compromise of an enterprise network. Two behaviour types, namely small payload beaconing (55%) and large payload egress, both using port 80, are captured in a total of 1821 sessions on 1 edge.
2. *IDS Beacons*: Traffic from the unfiltered backbone data weakly associated[3] with signatures of beaconing. 433 sessions on 6 edges.
3. *IDS Trojans*: Traffic from the unfiltered backbone data weakly associated with signatures of select trojans. 331 sessions on 6 edges.
4. *IDS Keep-Alives*: Traffic from the unfiltered backbone data weakly associated with signatures of select keep-alive events. 1115 sessions on 7 edges.

Seeded sessions account for 3740 out of 9 058 815 sessions, i.e. 0.04%, or approximately 1 in 2500.

[3] A weak association is made by filtering signature matches by either SrcIP or DstIP but not necessarily both.

4. Modelling with Netflow Literals

These tests propose simple features with which to check the suitability of DMM for modelling diverse Netflow metadata. That is, when the model is simple, will it converge, will the source distributions and related clusters provide useable insights into traffic composition, and will likelihood tests provide useful demarcations between hosts or classes of traffic?

4.1. Feature Selection

A *feature set* was chosen that provides a simple representation in each of the main categories of connectivity, size and timing:

- *Degree*: A measure of connectivity by out-degree, given as
  $D_{out} = \log_2(\#\text{ unique end-points per host}) \in [0,1, 2, \ldots, 12]$.
- *Packet Count*: A measure of size by the distribution of packet counts, given as
  $N_{pkt} = \{ \#\text{ packets per session} \} \in [1, 2, \ldots, 30]$.
- *Octet Count*: A measure of size by the distribution of byte counts, given as
  $N_{oct} = \{ \log_2(\#\text{ bytes per session}) \} \in [5, 6, \ldots, 17]$.
- *Span*: A measure of timing by the distribution of session durations, given as
  $T = \{ \log_2(T_{end} - T_{start} \text{ per session}) \} \in [-9, -8, \ldots, 11]$.

These are absolute value representations of the unaltered, or literal, Netflow fields. An *observation* takes the form of four concatenated distributions binned as above, as illustrated in Figure 5. Per-dimension, out-of-range values are attributed to the nearest bin-end.

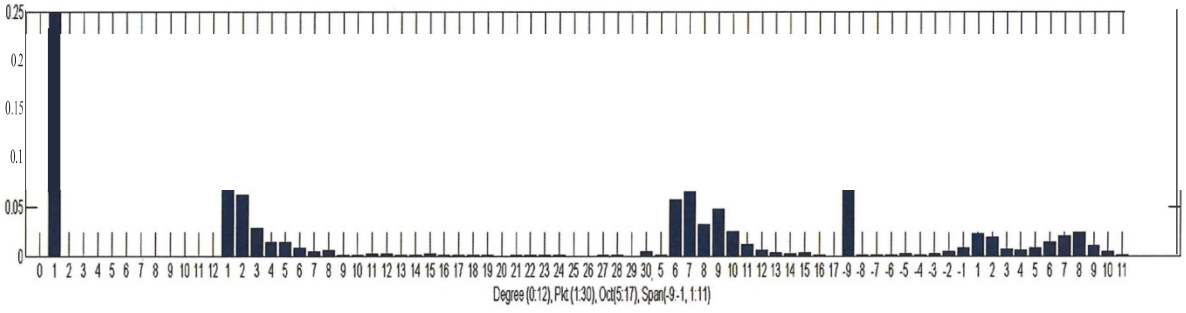

Figure 5: A single observation using Netflow literals.

An aggregation strategy for *observation construction* is also required. Here aggregation is *per-host*, i.e. all sessions per host (IP) are accrued to build the selected distributions. Although by the client-server model of internet communications [7] each Netflow record should show the client as the originator of the session (i.e. on the source side), imperfect collection frequently allows the opposite. Since the behaviours of a client might reasonably be expected to be markedly different from those of a server, a more insightful and consistent model is likely if the client- and server-sides are profiled separately. Here the

server-side is modelled, which facilitates explanation with respect to the selected class (service) labels, and all sessions are remapped such that the server is (arbitrarily) on the source-side of the record, based on the (imperfect) assumption that the lowest port value in each session is usually associated with the server.

Note that the bins in each dimension here were chosen empirically. These could be established more systematically, for example by using an entropy measure such as relative uncertainty (RU, [17]) to ensure a robust representation (high diversity) as class intervals are iteratively made progressively narrower.

4.2. Modelling

Model parameters (Section 1.4) include X, θ, Z, ϕ, α and β, of which X, α and β must be given. The dimension of X is given by the observation construction above and also sets a dimension in θ and α. The dimension of β is arbitrary and reflects an estimation of the number of distinct sources (behaviours) expected. Once set, this fixes the sizes of θ and α. In this work the number of sources S was set at 32, and both α and β were chosen as uniform priors, i.e. set to unity. Hence X, α and β are known.

Gibbs MCMC solving follows initialisation of the other parameters, and was limited to 1000 iterations. Convergence was judged using the source proportion estimates ϕ, as illustrated in Figure 6.

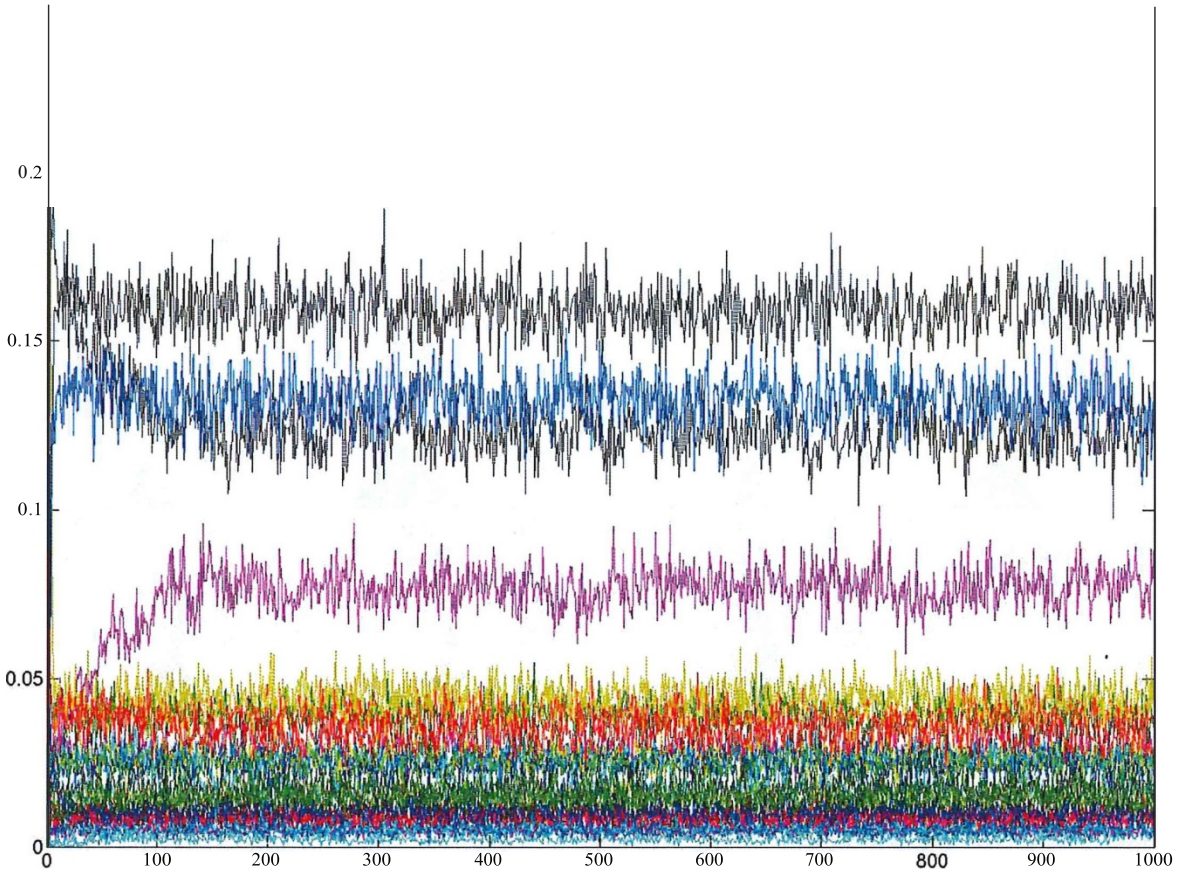

Figure 6: Example of source proportion convergence during MCMC solving. A coloured trace represents (arbitrarily) the estimated occupancy of a source from S, and reaches a stable mean.

Each element of parameter vectors θ, ϕ and Z accrues an estimated distribution with 1000 samples. For example, the estimates of source proportions are shown in Figure 7.

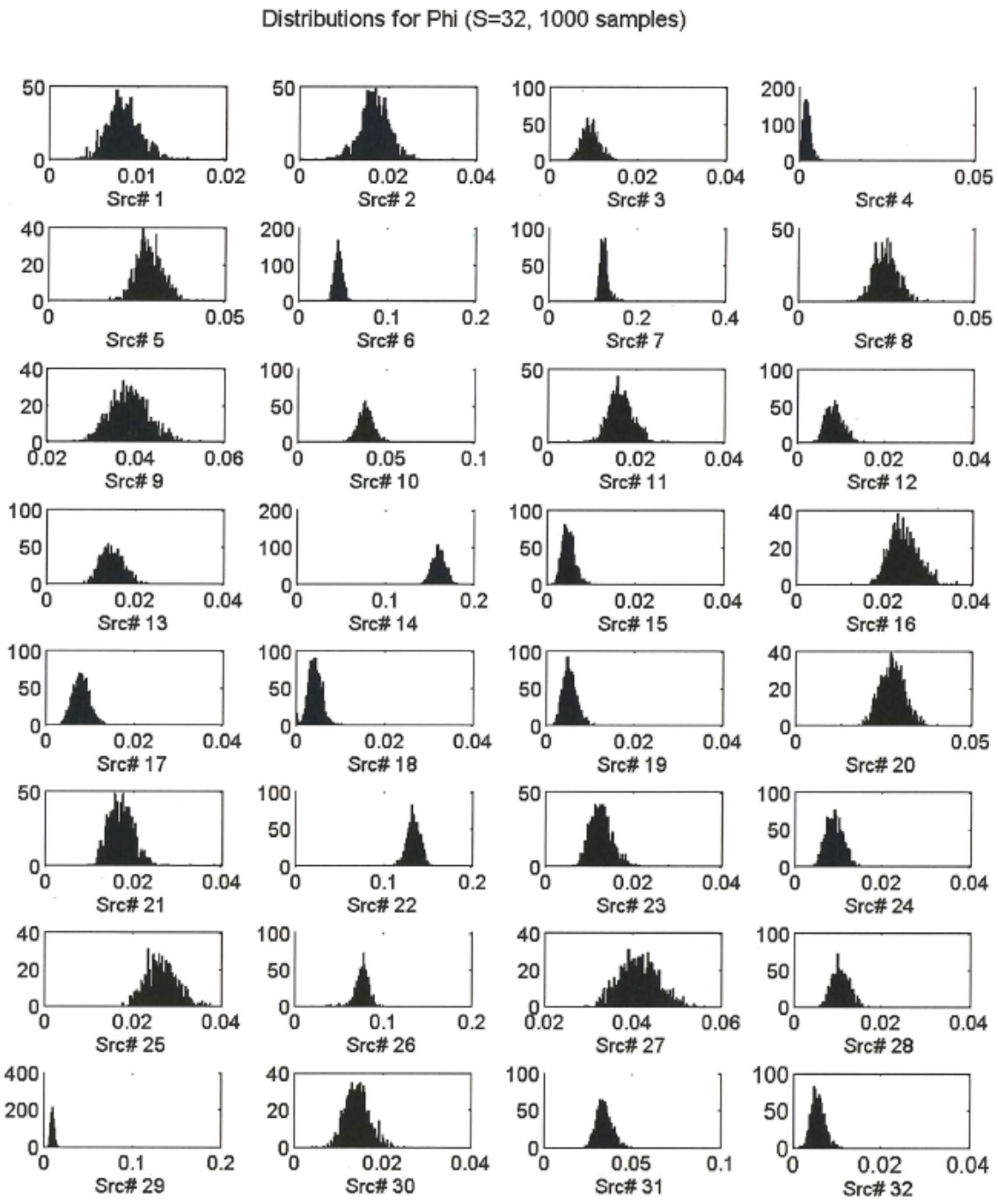

Figure 7: Example distributions for S source proportion estimates. The x-axes are absolute values and the y-axes the counts in observed class intervals during iterative estimation.

Where point estimates of the values are required (for example to define an instance of the model with which to calculate likelihoods), these were taken simply as the modes of the distributions (the *maximum a-priori* or MAP values).

With 2196 observations constructed according to degree, packet count, octet count and span the model suggests (at least) 32 different behaviours described further in the following sections.

4.3. Cluster Utility

Clusters in the DMM solution are related to *behavioural* similarities in the feature set, as determined by the applied joint probability model and not necessarily related to a conventional distance metric [18]. The extent to which this may relate to traffic *labelling* is a function of whether the various classes so labelled are behaviourally distinct, and whether the label itself is a fully resolved feature. In this experiment the class label (the service port) was not an explicit feature, and there is no expectation that the clustering solution should neatly differentiate traffic by class. Rather, there could be multiple clusters of the same class if it exhibits distinct behaviours, or clusters with multiple classes if they exhibit similar behaviours. The utility of the cluster solution must be assessed on this basis.

Observations were clustered into 32 sources, described by general attributes in Table 3. For each source:

- The *Src* # column identifies non-empty sources.
- The *Host Count* is the number of SrcIP considered members of the cluster. This may differ from the host counts in Table 2 (selected from heavy-hitters in each port class, Section 3.2) because the source-as-service mapping creates 'new' hosts where the original filter constraint was a DstPt but the SrcPt nonetheless has a smaller value (or vice versa).
- *Seed Count* indicates the number of seeded edges in the cluster.
- *Port Tendency* reveals the (top-3) dominant SrcPt in each cluster, which may relate to Class.
- *Behavioural Tendency* is a subjective description of trends in each cluster based on feature distribution quartiles.
    - Degree may be low, medium, high or very high.
    - Packet-count and octet-count may be low, medium, high or diverse (present across all quartiles).
    - Span may similarly be short, medium, long or diverse.

There are several examples where the appearance of members of the same traffic class in different clusters can be related to behavioural distinctions. For example, HTTP(S) services dominate clusters 2, 4, 11, 24 and 28. The behaviour in 2, where there are large client bases and full utilisation of the ranges for payload size and duration, appears typical of legitimate and popular web servers. 28 is similar but with reduced scale in client base,

payload size and duration. 4 and 24 however display point-to-point connectivity and have consistent (small) payloads, with 4 having long duration sessions and 24 having short. This ability to segregate less common behaviours in standard protocols could be useful in a detection and filtering context.

Similarly, clusters of RDP (3, 8, 17 and 20) are differentiated largely by out-degree and clusters of ICMP (6, 7, 10, 14, 22 and 26) are differentiated largely by size, allowing a segregation of *echo request* (6) from *port unreachable* types, and showing observations from several other protocols as similar. Most members of these large ICMP clusters are in fact a consequence of the source-as-service mapping explained in Section 4.1.

Clusters showing similarity between observations from multiple classes include 1, 5, 12, 16, 23, 27 and 32. These can expose unexpected relationships, such as DNS in 16 appearing with an unusually large payload.

In terms of the charters of characterisation, detection and filtering the DMM cluster solution can contribute as follows:

- *Characterisation:* Identification of the number of distinct behaviours and their proportions.
- *Detection:* By inspection, the potential to identify atypical use of common services.
- *Filtering:* The ability to attribute an observation from the network to a source (behaviour class) that may be of interest.

Variations in feature selection and observation construction may produce entirely different results - this is explored in subsequent sections.

Table 3: Observation Clusters with the Netflow Literals Model

Src #	Host Count	Seed Count	Port (Class) Tendency	Behavioural Tendency			
				Degree	*Pkt Cnt*	*Oct Cnt*	*Span*
1	6	5	37% 1935 16% 38101 16% 23	Low	High	High	Long
				Point-to-point transactions from atypical server ports with large payload sizes.			
2	38		46% 80 43% 443	Very High	Diverse	Diverse	Diverse
				HTTP(S) services to large client sets with a wide variety of content types.			
3	19		97% 3389	Low/Med	Low/Med	Low /Med	Short/Med
				Remote Desktop Protocol services, with two distinct payload modes, being approximately 70% small and 30% medium (in size/ duration).			
4	3	1	97% 80	Low	Low/Med	Low/Med	Long

Src #	Host Count	Seed Count	Port (Class) Tendency	Behavioural Tendency			
				Degree	*Pkt Cnt*	*Oct Cnt*	*Span*
				Point-to-point HTTP with bimodal (50/50) size but long duration.			
5	62		55% 53 15% 1024	Med	Low	Low	Short
				Potentially DNS services, but with a small payload size.			
6	99		100% 0	Low	Low	Med	Short
				ICMP *(Echo Request)*, typically with packet size of the order of 1 kB			
7	279		75% 0 22% 3074	Low	Low/High	Low/High	Short/Long
				ICMP *(Port Unreachable)* with bimodal size and duration in proportion to the ports.			
8	55		80% 3389	Low	Low	Low	Short
				RDP with consistently low size, and a limited set of ephemeral destination ports.			
9	85	2	80% 8080	Low	Med	Med	Med
				Proxy HTTP with a variety of mid-range sizes and durations.			
10	85		59% 0 21% 5060	Low	Low	Med	Short
				ICMP *(Port Unreachable)* and SIP with consistently low packet count and a variety of sizes up to lkB.			
11	36		47% 80 35% 1433	Med	Med	Med	Med
				HTTP and MS SQL with similar payload sizes.			
12	18		29% 23 22% 6969 18% 5242	High	Low /Med	Low /Med	Med
				Telnet and possibly bittorrent services to many clients, with an even spread of packet and byte counts up to medium size, and bimodal span.			
13	32		93% 1433	Med	Low	Low	Med
				MS SQL services to moderate numbers of clients, with small sizes.			
14	359		67% 0 33% 3074	Low	Low /High	Low /High	Short/Long
				ICMP *(Port Unreachable)* with bimodal size and duration in proportion to the ports.			
15	10		71% 6881	Very High	Low /Med	Low /Med	Long

Src #	Host Count	Seed Count	Port (Class) Tendency	Behavioural Tendency			
				Degree	*Pkt Cnt*	*Oct Cnt*	*Span*
			25% 53	Bittorrent services to many clients, with diverse small sizes and long duration.			
16	53		45% 53 14% 12350	Very High	Low..High	Low..High	Short..Long
				DNS and possibly Skype services to many clients, with diverse size and duration distributed exponentially.			
17	*17*		88% 3389	Low	Low	Low	Short
				RDP with consistently low size.			
18	9		91% 3306	Low	Med	Med	Med
				Point-to-point MySQL connections.			
19	11		89% 445	Very High	Low	Low	Med
				MS DS with many clients and consistent size, duration.			
20	60		78% 3389	Low	Low	Low	Short
				RDP with consistently low size.			
21	37		64% 53 25% 6881	Very High	Low	Low	Short
				DNS and bittorrent services to many clients, with consistent size and duration.			
22	296		90% 0	Low	Low	Low	Short
				ICMP *(Port Unreachable)* with consistently small size.			
23	27		31% 8883 31% 3072 17% 5222	Med	Low	Low	Short
				A mix of services including Facebook with typically small size and duration.			
24	19		71% 80 13% 1024	Low	Low	Low	Short
				Point-to-point HTTP with consistently small size.			
25	58		45% 5223 25% 53	Very High	Low..High	Low..High	Short..Long
				XMPP to many clients with exponentially distributed size and span.			
26	172		65% 0	Low	Low	Low	Short
			18% 5060	ICMP *(Port Unreachable)* with consistently small size.			
27	90		19% 4244 12% 3072 11% 6890	Very High	Low..High	Low..High	Short..Long
				Possibly bittorrent or Viber desktop services to many clients, with exponentially distributed size and span.			

Src #	Host Count	Seed Count	Port (Class) Tendency	Behavioural Tendency			
				Degree	*Pkt Cnt*	*Oct Cnt*	*Span*
28	23		94% 80	Med	Med	Med	Med
				HTTP with moderate size and durations.			
29	22	2	54% 8080	Low	Low	Low	Med
				Proxy HTTP with consistently small size and duration.			
30	31	10	56% 23	Low	Med/High	Med	Med
			12% 110 9% 25	Point-to-point telnet and mail services.			
31	73	1	69% 110	Low	Low /Med	Low /Med	Short/Med
				Point-to-point POP3 mail.			
32	12		54% 8080 15% 3074 14% 5060	Med	Low /Med	Low /Med	Short/Med
				Mixed services with bimodal distributions in size and span.			

4.4. Likelihood Utility

According to the model, each observation (in this case, an IP address associated with a server) has a numeric likelihood of *belonging* to the model (strictly, the probability of the sample given the prior observations, the values of the priors, and the estimates of the source proportions and distributions). Ignoring the *evidence* in the Bayesian model, a relative probability may be easily computed and is typically expressed in base-10 logarithmic form as the log-likelihood, or *loglik*.

Whereas the source allocations from the model allow the question 'What is similar?', the 1ogliks allow interrogations of the form 'What is (a)typical?' . This work provides the following analysis based on logliks:

- Identification of the (say, 10) least-likely and most-likely hosts.
- Ascertaining whether the seeded traffic may be differentiated by likelihood.
- Checking whether the distribution of logliks is modal, and if so whether the modes resolve to distinct classes or behaviours.
- Re-computing host likelihood across a time-series of smaller observation windows to ask whether the distribution of loglik per-host forms a more complex behaviour vector.

4.4.1. Likely and Unlikely Observations

The most likely observations (Table 4) all relate to ICMP transactions with small payloads and short duration. The least likely (Table 5) are less common services showing a bias to large, slow payloads.

Table 4: The 10 Most-Likely Observations in the Netflow Literals Model

Rank	# Dst	SrcPt	DstPt	Behavioural Tendency			
				Degree	*Pkt Cnt*	*Oct Cnt*	*Span*
1..10	1	0	771	Low	Low	Low	Short
				All of the most likely observations are hosts with low session counts engaged in ICMP with a single destination, where the sessions have one packet, 64 bytes and nominally zero-span.			

Table 5: The 10 Least-Likely Observations in the Netflow Literals Model

Rank	# Dst	SrcPt	DstPt	Behavioural Tendency			
				Degree	*Pkt Cnt*	*Oct Cnt*	*Span*
2196, 2193, 2192, 2191	9	0	2048	Low	High	High	Long
				Prominent among the least likely are ICMP (Echo Request) where the payload is large and the sessions are slow.			
2195, 2194	681, 524	3074	Ephemeral	Very	High	High	Long
				XBOX service.			
2190	235	5222	Ephemeral	High	Low	Med	Short
				XMPP service.			
2189, 2188	2	5060	Ephemeral	Low	Low/Med	Low/Med	Short/Long
				SIP service.			
2187	54	10000	10000	Med	Low/High	Low/Med	Short/Long
				Unknown protocol with bimodal payload size and duration.			

4.4.2. The Loglik Distribution

Figure 8 provides a three-part description of the overall distribution of likelihoods for the full set of (2196) observations. Top is the distribution from all observations - a largely uniform (uninformative) spread of values, except where interspersed with several

significant peaks. Middle is the subset of likelihoods associated with only the seeded observations. Again, these are spread widely across the observed range of outcomes, suggesting that a point-measurement of likelihood alone is not a useful discriminator of their behaviours. This is reinforced by the bottom part, which shows the cumulative proportion of seeded (red) versus non-seeded (blue) observations as they are counted in ascending order of likelihood -no useful demarcation is provided.

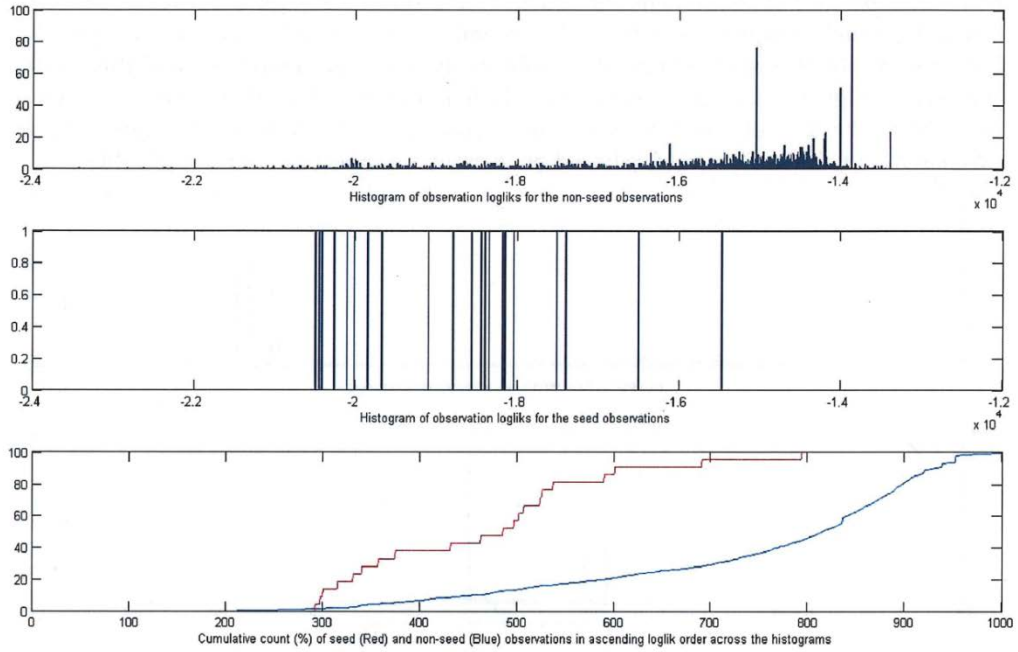

Figure 8: Distributions of likelihoods in the Netflow Literals model

It remains to check the consistency of observations in the peaks. Consider the three largest peaks near the relative logliks -15000, -14000 and -13900:

- -15000: 74 hosts using ICMP (Echo Request) with single-packet, 1kB, short-duration sessions. This aligns closely with cluster 6 in Section 4.3.
- -14000: 51 hosts using ICMP (Port Unreachable) with single-packet, 64B, short-duration sessions. This aligns closely with cluster 22 in Section 4.3.
- -13900: 85 hosts using ICMP (Port Unreachable) with single-packet, 64B, short-duration sessions. This aligns closely with cluster 22 in Section 4.3.

Hence there is limited potential for modality in the likelihood distribution to inform behavioural segregations.

4.4.3. Host Behaviour as a Vector of Likelihoods

Although peaks in the loglik distribution could be related to particular traffic types and behaviours, overall there was limited modality with which to characterise, detect or filter

the bulk of the observations. This could be partly a function' of aggregation - by deriving each observation from all of a host's traffic, behaviour over smaller time windows that could be discriminating is potentially suppressed. To check whether the *pattern of behaviours* associated with a host has relevance, an alternate set of observations based on short session aggregates was computed and the likelihood's of these samples checked.

First, Figure 9 shows the equivalent of Figure 8 above, for the revised set of likelihoods. Note that the distribution remains dispersed, although the likelihoods of the seed hosts are grouped with substantially less variance. Secondly a new *likelihood distribution per hos*t measurement constructed from the set of logliks per host is shown in Figure 10, which samples the first and last members as ranked by count of occupied bins. Only a minority of hosts (20%) display multiple 'behaviours' (Figure 11) and this includes most seeds (Figure 12). Such observations allude to potential for *filtering* by this measure - for example the '20% of hosts' threshold also demarcates 50% of the seeds (Figure 11). In a *detection* role, a similarity measure based on vectors of interest might be applicable.

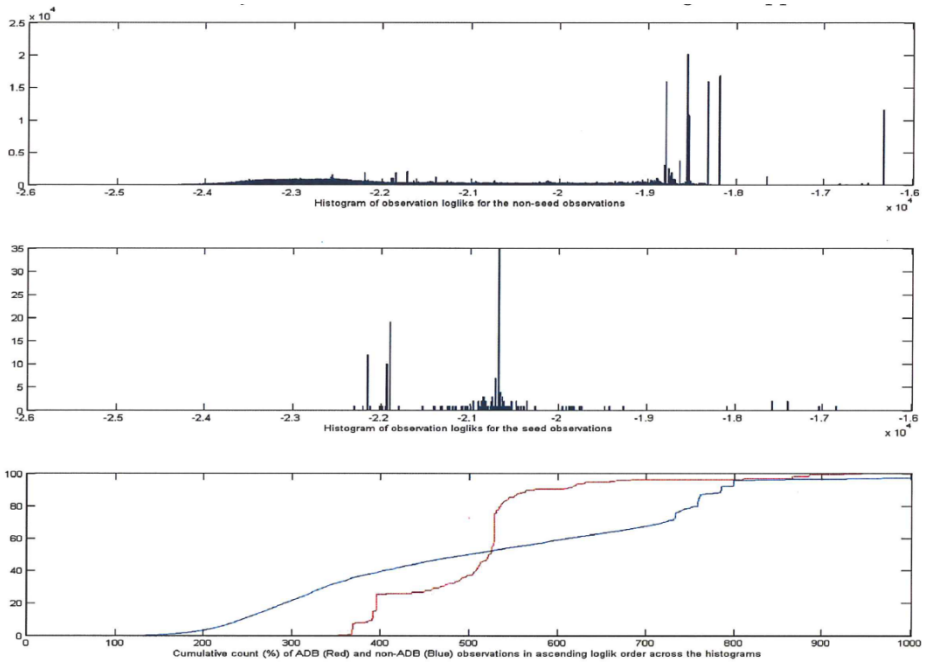

Figure 9 - Distributions of likelihoods in the Windowed Literals model

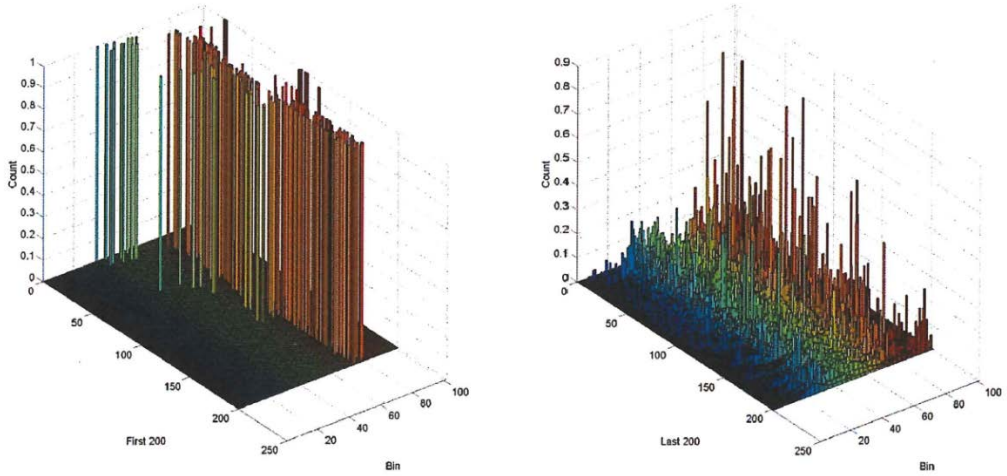

Figure 10 - Examples of likelihood-per-host vectors, ranked by bin-count

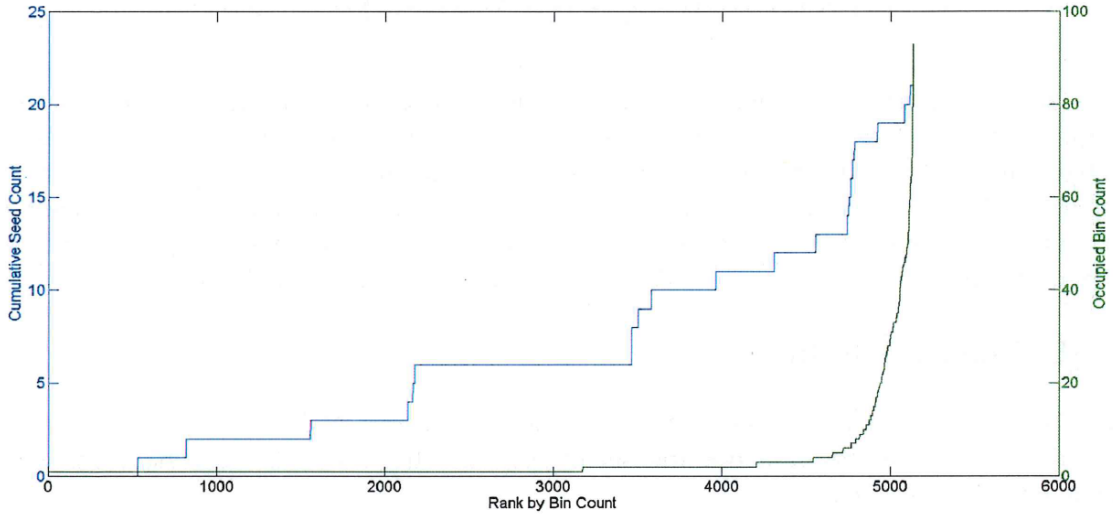

Figure 11 - Ranked bin-count in likelihood-per-host vectors (green), with cumulative sum of seeds (blue)

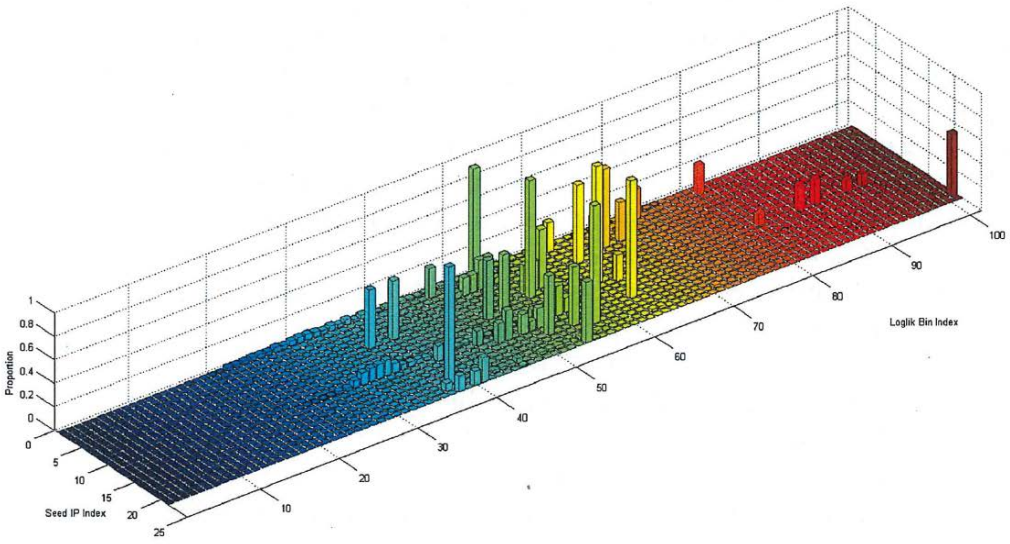

Figure 12 - Likelihood-per-host vectors for the seeds

4.5. Summary

The clustering aspect of a DMM model can reveal behavioural segregations in traffic, supporting source characterisation and limited detection of anomalies by inspection. However, as a function of aggregation, the likelihood solution can be uninformative. Aggregation to yield a single measurement of likelihood per host was uninformative in this case. Splitting each host's traffic into smaller time-ordered aggregates and hence allowing multiple measurements of likelihood per host provides additional profiling opportunities.

5. Modelling with Heuristic Features

The behavioural resolution of the modelling in Section 4 is limited by the choices of aggregation and feature vector. By aggregating observations on a per-host basis, distinct behaviours within a host's traffic are potentially obscured. By limiting the features to representations of absolute value, behaviours related to (at least) sequential properties are excluded. For comparative purposes, tests in this section retain per-host aggregation, but expand the feature vector to permit detection of more complicated behaviours.

5.1. Feature Selection

There is a speculative element to the selection of features. Neither the set of behaviours actually present in the data nor the set of features that allow those behaviours to be resolved are known in advance.

Considering the suggested feature classes of identity, connectivity, size and timing, the previous model ignored identity (the destination IP addresses), addressed connectivity by out-degree, size by packet and byte counts, and timing by span, all as absolute values. A more generic feature vector would include identity, could better describe connectivity by the sets of source and destination ports used, and could use not only absolute value but also representations of, say, *differences* between values to elucidate sequential behaviours, and *self-similarity* [19] to elucidate more complex patterns of change or repetition. According to this philosophy the following features were proposed:

- The histograms of the *absolute* values, aggregated per-host, for the following dimensions and their respective class intervals:
    - 1st Octet ... 4th Octet of DstIP ε [0, 8, 16, 24, ..., 248] (32 bins).
    - SrcPt ε [$2^{(0, 0.5, 1, 1.5, ..., 16)}$] (33 bins).
    - DstPt ε [$2^{(0, 0.5, 1, 1.5, ..., 16)}$] (33 bins).
    - Packet-count ε [1, 2, 3, ..., 30] (30 bins).
    - Byte-count ε [$2^{(5, 6, 7, ..., 17)}$] (13 bins).
    - Span ε [$2^{(-9, -8, ..., 10)}$] (20 bins).
- The histograms of the *differences* of the same set of dimensions, taken from the absolute values of change between adjacent elements in the time-ordered sets, and over the same sets of class intervals.
- The histograms of the *self-similarity* scores of the same set of dimensions, taken from the upper-triangular part of the matrix of absolute values of differences between all elements in the set (limited to 1000 members chosen randomly for computational efficacy if the set contained more than 1000 members), normalised and binned in range [0, 0.05, 0.1, ..., 0.95] (20 bins).

Again, out-of-range values are attributed to the nearest bin-end.

The aggregation strategy used in Section 4.1 was retained, and for each host an observation formed by normalising and concatenating the given 27 histograms per host.

5.2. Modelling

Modelling follows the approach set forth in Section 4.2. Note that with the selected features, each observation has 694 bins, and in the MCMC solving a Dirichlet distribution of the same dimensionality must be sampled. Under these conditions convergence was not occurring in a short time frame. This was solved by sampling each of the 27 dimensional subsets of the observations separately, and the model converged with 3 apparent sources.

5.3. Cluster Utility

Although the model building process successfully converged, the source estimates comprised only three distinct sources: two ICMP differentiated mainly by the destination addresses, and one being everything else (Table 6). Accordingly, these results have little utility in characterising the source.

Table 6 - Observation Clusters with the Heuristic Model

Src #	Host Count	Seed Count	Port (Class) Tendency	Behavioural Tendency			
				Degree	*Pkt Cnt*	*Oct Cnt*	*Span*
2	353		100% 0	Low	Low	Low	Short
				ICMP *(Port Unreachable)* with consistently small size.			
16	465		100% 0	Low	Low	Low	Short
				ICMP *(Port Unreachable)* with consistently small size.			
27	1378	21	33% 80 29% 443 12% 53	High	Diverse	Diverse	Diverse
				The majority of hosts clustered together with highly diverse traffic.			

5.4. Likelihood Utility

It remains of interest to establish the utility of the likelihoods when the source model is poor; hence the analysis from Section 4.4 is repeated here for this model variant.

5.4.1. Likely and Unlikely Observations

Consistent with the previous model the most likely observations (Table 7) are largely sets of short sessions of ICMP traffic, although alternate services with similar small payload and point-to-point attributes also emerge. The least likely (Table 8) are web and mail servers -an unintuitive result and a consequence of the limited model resolution.

Table 7 - The 10 Most-Likely Observations in the Heuristic Model

Rank	*# Dst*	*SrcPt*	*DstPt*	*Behavioural Tendency*			
				Degree	*Pkt Cnt*	*Oct Cnt*	*Span*
1	1	8000	Ephemeral	Low	Low	Low	Short
				Possibly Internet Radio.			
2	1	6881	Ephemeral	Low	Low	Med	Short
				Bittorrent with moderate packet size.			
3	1	5060	5060	Low	Low	Med	Short
				SIP protocol.			
4..10	1	0	771	Low	Low	Low	Short
				A majority of the most likely observations are hosts with low session counts engaged in ICMP with a single destination, where the sessions have one packet, 64 bytes and nominally zero-span.			

Table 8 - The 10 Least-Likely Observations in the Heuristic Model

Rank	# Dst	SrcPt	DstPt	Behavioural Tendency			
				Degree	*Pkt Cnt*	*Oct Cnt*	*Span*
2196, 2195, 2193, 2191, 2187	O(10000)	80/443	Ephemeral	High	Diverse	Diverse	Diverse
				Servers of HTTP with wide out-degree and a diverse range of content.			
2194	1644	1024	Ephemeral	High	Diverse	Diverse	Diverse
				An unusual service port, to a wide range of hosts with a wide range of content.			
2192, 2190, 2189, 2188	O(10000)	25	Ephemeral	High	Diverse	Diverse	Diverse
				Mail service to a wide range of hosts with a wide range of content.			

5.4.2. The Loglik Distribution

Figure 13 provides the three-part description of the likelihoods. Compared to Section 4.4.2, the overall distribution (top) has more structure, but the seeds are again evenly distributed through the observations (middle and bottom), making point measurement of likelihood a poor detector.

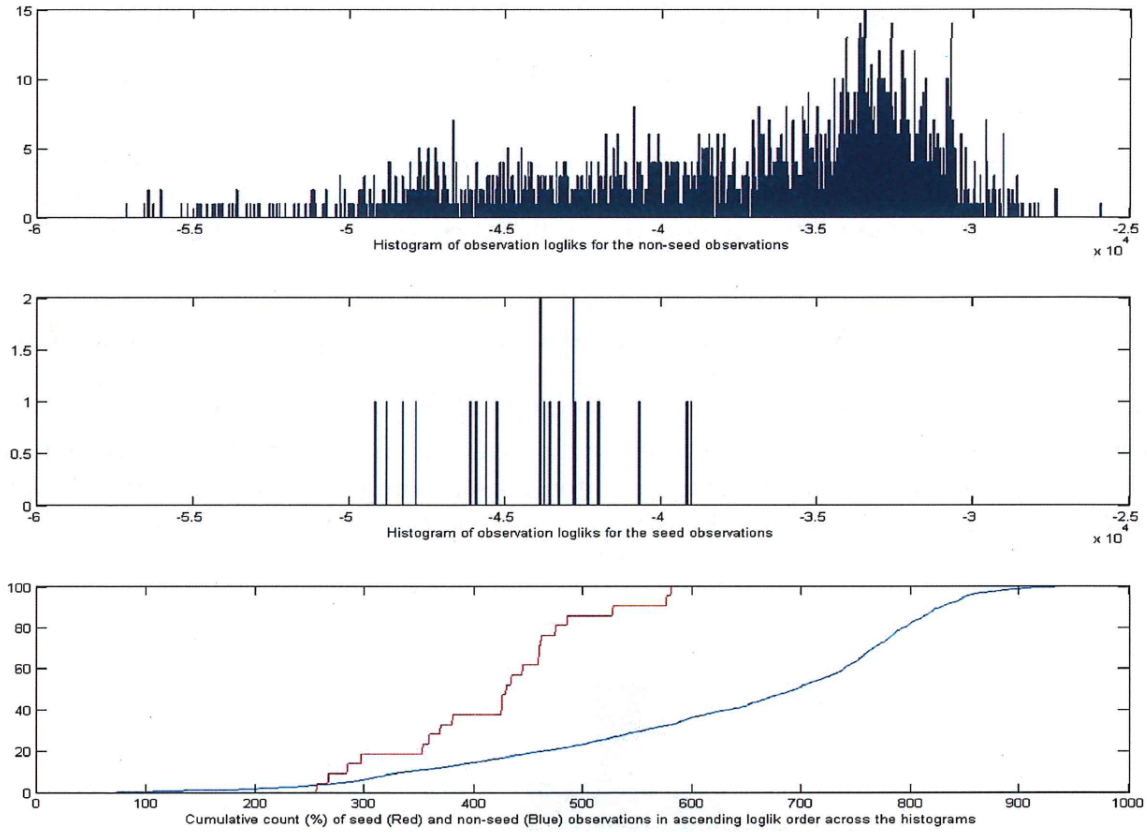

Figure 13 - Distributions of Likelihoods in the Heuristic Model

Checking for behavioural structure in the modes of the distribution:

- The region of likelihood -37000 to -32000 shows a higher density of RDP (3389) hosts.
- The region of likelihood -42000 to -39000 shows a higher density of XMPP (5223) hosts.
- The tail below -42000 has a service distribution consistent with the overall traffic distribution, i.e. provides negligible demarcation.

There is limited potential for modality in the likelihood distribution to inform behavioural segregations, consistent with the limited source model.

5.5. Summary

With the chosen features the DMM model collapses to a few clusters, providing very limited resolution of different behaviours. Counter-intuitive likelihood results follow, where the least-likely observations are selected from the most ordinary of traffic classes (i.e. web and mail serving).

The source model gives the impression that most traffic looks 'the same'. For the model to resolve different behaviours in the observations, the observations must be separable. Section 6 explores the separability of the observations based on correlation coefficients for the two feature vectors considered thus far, then proposes and assesses an alternative feature selection process.

6. Modelling with Minimised Observation Correlation

Subjective assessment of dimensions that could contribute to behavioural segregations in Section 5 proved a poor choice for DMM modelling. The existence of a single source for the vast majority of the traffic suggests that with the chosen features, all traffic looks 'the same'.

The 'sameness' of the constructed observations could be quantified by a similarity measure, such as a correlation coefficient test [12]. Comparing each feature vector to each other feature vector, taking the upper-triangular part of the (symmetrical) correlation matrix to eliminate duplicates, and plotting the distribution of correlation coefficients so obtained gives the result shown in Figure 14 for (top) the simple features in Section 4.1 and (bottom) the complex features of Section 5.1. Note that the former has a greater tendency for lower correlation (more distinct observations).

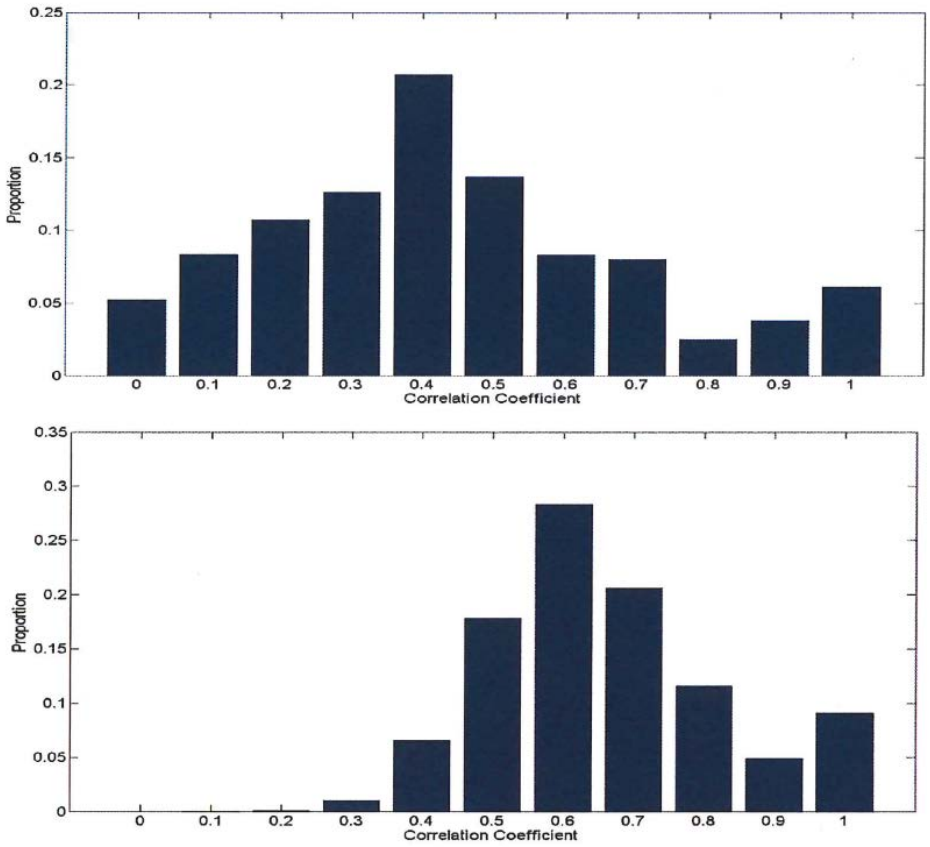

Figure 14 - Observation Correlation Scores for the Features of Sections 4.1 and 5.1

An improved methodology for feature selection could therefore be to propose features relevant to the behavioural segregations of interest from domain knowledge, then

moderate that selection by assessing all dimension combinations to find the subset with lowest correlation coefficient distribution (e.g. by taking the distribution mean).

6.1. Feature Selection

According to the lowest correlation paradigm, applied to the features in Section 5.1, the observations are most separable with the feature set comprising only the absolute value representations of the destination address and the byte counts:

- 1st Octet ... 4th Octet of DstIP ε [0, 8, 16, 24, ..., 248] (32 bins).
- Byte-count ε [2 (5, 6, 7, ···, 17)] (13 bins).

This subset has the correlation score distribution shown in Figure 15, which is notably improved compared to those above.

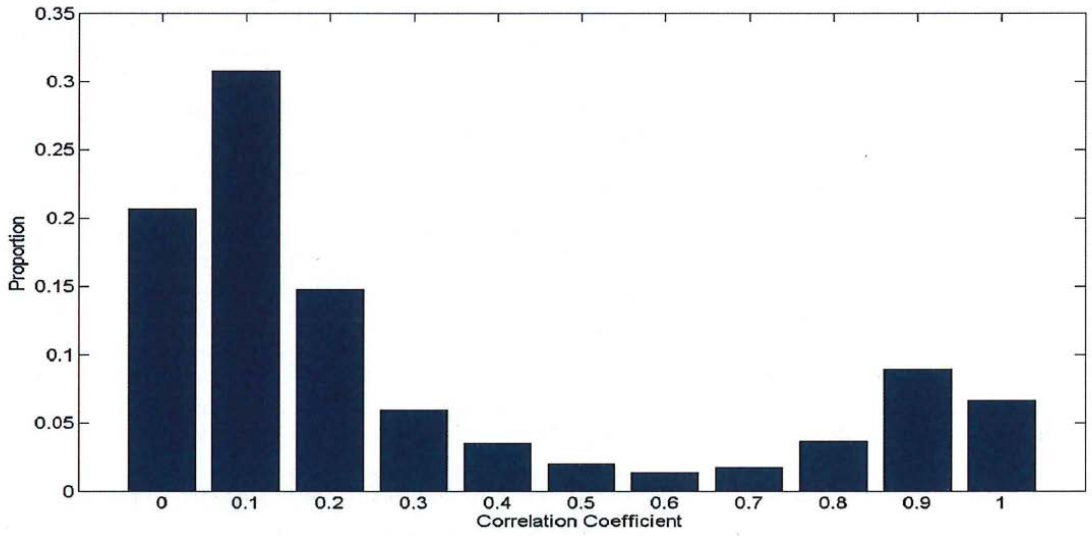

Figure 15 - Observation correlation scores for the Features of Section 6.1

The observations for modelling were the 2196 per-host aggregates based on this reduced feature set.

6.2. Modelling

The modelling strategy from Section 4.2 was retained and model convergence was achieved with 20 apparent sources.

6.3. Cluster Utility

The model resolves 20 clusters, with 11 dominated by a single class, of which 4 are ICMP variants and the others are unique (Table 9). Previously, the model based on simple Netflow literals resolved about 6 unique classes into distinct clusters, so overall the proportions are similar.

Again, there is a single large cluster with many similar traffic types and services lumped together, whereas the less common usages of a service appear to originate other clusters. Hence the gist of the contributions to characterisation, detection and filtering from this solution are similar to those assessed in Section 4.3.

Table 9 - Observation Clusters with the Optimised Model

Src #	Host Count	Seed Count	Port (Class) Tendency	Behavioural Tendency			
				Degree	*Pkt Cnt*	*Oct Cnt*	*Span*
2	5		99% 53	High	Low	Low	Short
				DNS with a consistently small size of the order of 64 bytes.			
3	15		97% 8080	Low	Low	Low	Med
				Proxy HTTP with constant size characteristics.			
4	24		95% 3389	Low	Low	Low	Short
				RDP with consistently low size.			
5	139		90% 445	Med	Low	Low	Med
				MS DS.			
6	14		53% 123 21% 3072	High	Low	Low/Med	Short
				NTP service.			
7	390		46% 80 31% 53	Very High	Diverse	Diverse	Diverse
				HTTP and DNS services to large client sets with a wide variety of content types.			
9	112		68% 53	Very High	Low	Low	Short
				DNS with a large client base and consistently small size of the order of 128 bytes.			
12	46	2	46% 1024 14% 1025 11% 1026	Med	Low /Med	Low/Med	Low/High
				Possibly bittorrent with exponentially distributed size and bimodal span.			
13	3		100% 0	Low	Low	Low	Short
				ICMP *(Echo Request)*.			
14	12		95% 6881	Low	Low	Low	Short
				Bittorrent with consistently small size.			
15	105	1	46% 0 24% 80	Low	Low	Low	Short/Long
				Mixed services to small client sets, with small sizes and short or long duration.			

Src #	Host Count	Seed Count	Port (Class) Tendency	Behavioural Tendency			
				Degree	*Pkt Cnt*	*Oct Cnt*	*Span*
			24% 3389				
16	4		62% 0 38% 5060	Low	Low	Low/Medium	Short/Long
				ICMP *(Port Unreachable)* and SIP with consistently small size			
19	1		100% 8000	Low	Med	Med	Med
				Possibly a streaming radio connection.			
22	1159	18	35% 443 34% 80	Very High	Diverse	Diverse	Diverse
				HTTP(S) services to large client sets with a wide variety of content types.			
23	96		99% 0	Low	Low	Low	Short
				ICMP *(Port Unreachable)* with consistently small size.			
26	37		64% 123 36% 0	Med	Low	Low	Short
				NTP and ICMP *(Port Unreachable)* to moderate client bases with consistently small size.			
27	1		100% 0	Low	Low	Low	Short
				ICMP *(Port Unreachable)* with consistently small size.			
28	25		47% 445 43% 5060	Med	Low	Low	Short/Long
				MS DS and SIP to moderate client bases with small size and bimodal duration.			
29	7		100% 0	Low	Low	Low	Short
				ICMP *(Port Unreachable)* with consistently small size.			
32	1		100% 12350	Very High	Low/Med	Low/Med	Short/Med
				Possibly the Skype service to many clients, with diverse size and duration distributed exponentially.			

6.4. Likelihood Utility

Likelihood tests are repeated to consider the effects of feature optimisation.

6.4.1. Likely and Unlikely Observations

The most likely observations (Table 10) are again ICMP. The least likely (Table 11) are mainly point-to-point mail transactions or scanning behaviour, noted with both remote connection and ICMP protocols. Of the three 'unlikely' sets produced thus far, the presence of scanning here makes this the most intuitively satisfying.

Table 10 - The 10 Most-Likely Observations in the Optimised Model

Rank	# Dst	SrcPt	DstPt	Behavioural Tendency			
				Degree	*Pkt Cnt*	*Oct Cnt*	*Span*
1..10	1	0	771	Low	Low	Low	Short
				All of the most likely observations are hosts with low session counts engaged in ICMP with a single destination, where the sessions have one packet, 64 bytes and nominally zero-span.			

Table 11 - The 10 Least-Likely Observations in the Optimised Model

Rank	# Dst	SrcPt	DstPt	Behavioural Tendency			
				Degree	*Pkt Cnt*	*Oct Cnt*	*Span*
2196	1	110	Ephemeral	Low	Low/Med	Low/Med	Short/Med
				Mail transactions to a single destination, with distinctly bimodal size and duration.			
2195	1	25	Ephemeral	Low	Low/High	Low/High	Short/Long
				Mail transactions to a single destination, with distinctly bimodal size and duration.			
2194	1	25	Ephemeral	Low	High	High	Long
				Mail transactions to a single destination, with generally high size and duration.			
2193	27	23	Ephemeral	Low	Low	Low	Long
				Telnet sessions to a small number of destinations with low payload size but long duration.			
2192	1	Ephemeral	Ephemeral	Low	Low	Low	Med
				Unspecified point-to-point sessions.			
2191	4578	6000	8089	High	Low	Low	Short
				Remote connection protocol to many hosts but with consistently small payload - possibly scanning.			
2190	15 560	0	2048	High	Low	Low	Short
				ICMP (echo-request) transactions from a Baidu crawler.			
2198	2	5060	Ephemeral	Low	High	High	Long
				SIP transactions with generally high size and duration.			
2188	9	5223	Ephemeral	Low	Med	Med	Med
				XMPP service.			
2187	1	110	Ephemeral	Low	Low	Low	Short
				Mail with small payloads.			

6.4.2. The Loglik Distribution

Figure 16 provides the three-part description of the likelihoods. Compared to both Section 4.4.2 and Section 5.4.2, the overall distribution (top) has substantially more structure, and the seeds are biased toward the lower quartile of the set (middle and bottom), making threshold filtering a potential mechanism for enriching seed-like behaviours.

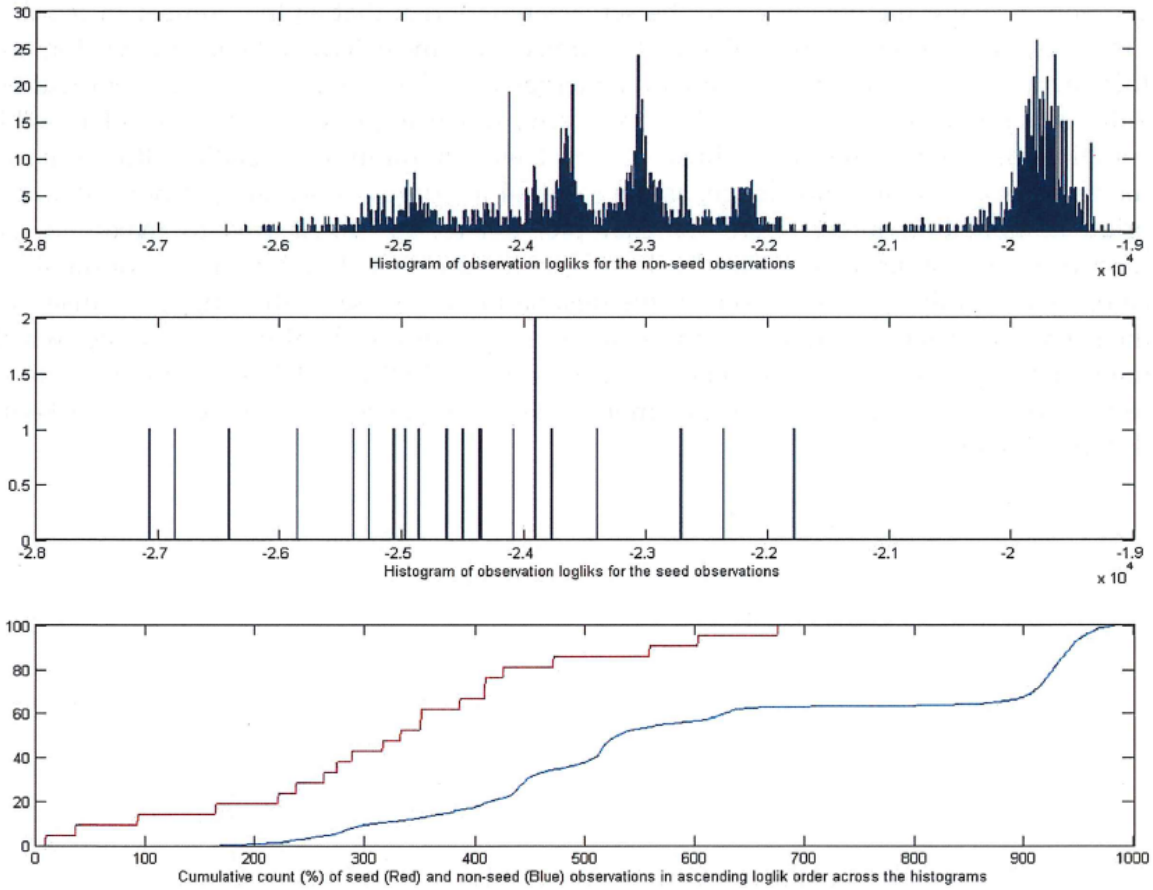

Figure 16 - Distributions of Likelihoods in the Heuristic Model

Checking for behavioural structure in the modes of the distribution:

- -22000 to -19500: 645 ICMP hosts, a subset of the main cluster #22.
- -22300 to -22100: 98 ICMP, MySQL and SIP hosts (60, 15 and 20% respectively).
- -22700 to -22600: 24 hosts using mail (40%) and NTP (35%).
- -23200 to -22900: 295 hosts with bias toward ports 1024 ... 1027-possibly bittorrent.
- -23750 to -23500: 244 hosts, mainly web servers.
- -24135: 20 hosts using X11PP (33%), HTTP proxy (25%) and bittorrent (20%).

- The tail below -24500 has 262 hosts with a service distribution consistent with the overall traffic distribution, i.e. provides negligible demarcation.

Excluding the tail, all modes have from one to a few dominant class members. Hence there is good potential for modality in the likelihood distribution to inform behavioural segregations. This could extend to better outcomes from profiling hosts by their vector of accumulated likelihood tests (Section 4.4.3).

6.5. Summary

DMM model utility presents as a function of the feature vector, which is heuristically related to the modelling objective. A simple representation from the main classes of information was sufficient to resolve a finite set of behaviours, from which aspects of characterisation, detection and filtering may follow, especially if the modelling objective presents in a specific behaviour in the set. There is a risk that as the number of features increases, more observations will look the same over some extent of the feature vector, and this increased correlation will lead to a merger of behaviours rather than segregation. Although this can be managed algorithmically, for example by examining all possible combinations of features to establish a subset with minimum correlation, this must be moderated by domain knowledge, as the residual feature set may suggest other than the original modelling intent. Here for example, where the intent was to maximise the discovery of distinct behaviours by having a feature-rich description, the optimal set comprised mainly representations of the destination addresses - this suggests clustering the servers nodes by similarity in the absolute ranges of the client base addresses, which might be useful for grouping related services like HTTP and DNS, or finding sets of servers that relate to specific autonomous systems or geographic zones - a markedly different intent.

7. Modelling Domains

In Sections 4 through 6 the models were built from traffic comprising both the heavy-hitter Netflow and the seeds. It is possible to build a model from only the seed (or non-seed) traffic to test whether this better segregates the two types by likelihood. It is also possible to build a model from a single class of traffic, and test whether this allows that class of traffic to be well segregated from others, and whether structure within that class can be elucidated. These test cases are considered below, using the base feature vector from Section 4.1.

7.1. Likelihoods from a Seed-only Model

A DMM model is converged from observations restricted to the seed-related Netflow according to the same methodology set forth in Section 4.2. There were 21 such observations. Then that model is used to test the likelihood of the full set of observations from prior tests. Figure 17 shows the comparative likelihood results. Note that the seeds are now the most-likely of the observations, which is intuitively satisfying. Further they are segregated from the majority of other observations by a gap in the likelihood histogram, which makes filtering decisions feasible - at the point where 90% of the non-seeds have been observed, only 10% of the seeds have been counted.

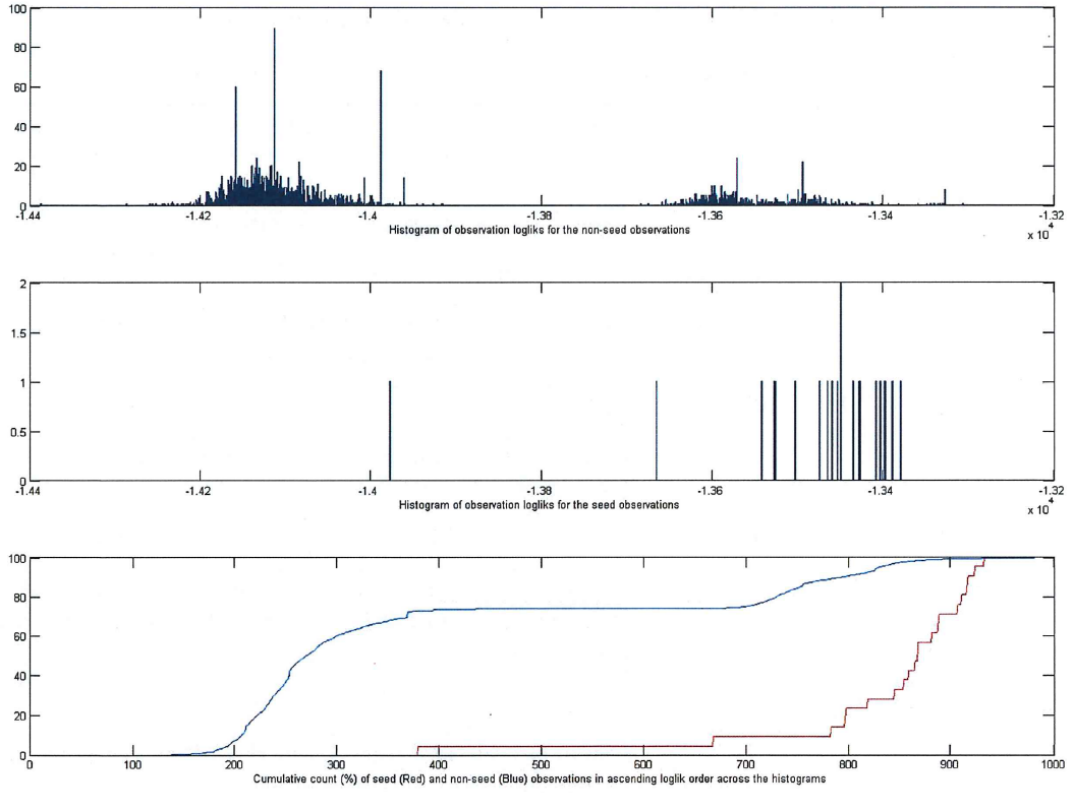

Figure 17 - Distributions of Likelihoods in the Seed-only Model

7.2. Likelihoods from an HTTP(S)-only Model

A DMM model is converged from observations restricted to the web-related (ports 80 and 443) Netflow according to the same methodology set forth in Section 4.2. There were 33 such observations, each corresponding to a web server in the dataset.

The source distributions for these observations resolve into eight clusters, described in Table 12. It is noteworthy that:

- The majority of HTTPS sessions occur in a single cluster (#5) with 26% of the HTTP. This suggests that the secure sessions are often distinct from ordinary web traffic.
- The majority of 'ordinary' HTTP sessions (with payload variety and large client sets) occur in just two clusters, #5 (30%) and #31(33%).
- Other clusters of HTTP traffic have distinguishing behaviours, including: # 2, 21 (a small client set where sessions have a long duration for a small payload - 5_ of the 6 hosts here are seeded); #6, 9 (bounded mid-range sizes and durations); #26 (many sessions with a single packet and 40-byte payload); and #30 (a mix of traffic but with a distinct mode from many sessions with a single packet and 44-byte payload).

The ability of the modelling to resolve such subclasses of behaviour could have useful application in detecting anomalous or unexpected traffic.

Table 12 - Observation Clusters with the HTTP(S) Model

Src #	Host Count	Seed Count	Port (Class) Tendency	Behavioural Tendency			
				Degree	*Pkt Cnt*	*Oct Cnt*	*Span*
2	2	1	100% 80	Low	Low	Low	Med /Long
				Small payloads, but long durations.			
5	7		74% 443 26% 80	Very High	Diverse	Diverse	Diverse
				Popular servers with a wide range of traffic types, largely using secure protocols.			
6	2	1	100% 80	Med	Med	Med	Med
				Payloads characterised by mid-range size and duration.			
9	2		100% 80	Very High	Med	Med	Med
				Popular servers with a range of traffic types distributed about the mid-range of payload size and duration.			
21	4	4	100% 80	Low	Low	Low	Long
				Small payloads, but long durations. Similar to cluster # 2.			
26	4		100% 80	Very High	Low	Low	Short
				Popular servers with a constant 1-packet, 40-byte payload.			
30	4		63% 80 37% 443	High	Low/Div.	Low/Div.	Low/Div.
				Popular servers with diverse payloads but a distinct repetition of low size and duration sessions.			
31	8		96% 80	Very High	Diverse	Diverse	Diverse
				Popular servers with a wide range of traffic types, largely using secure protocols.			

By likelihood ranking, the least likely servers were those with small payload but long duration sessions, including the SANS seed (Section 3.3). The most likely were the servers with constant (1-packet, 40-byte) sessions, continuing the trend seen with previous models.

The likelihood distribution for this small observation set is largely uninformative, but shown in Figure 18 for completeness. HTTPS traffic is demarcated below the null near minus l.32x104.

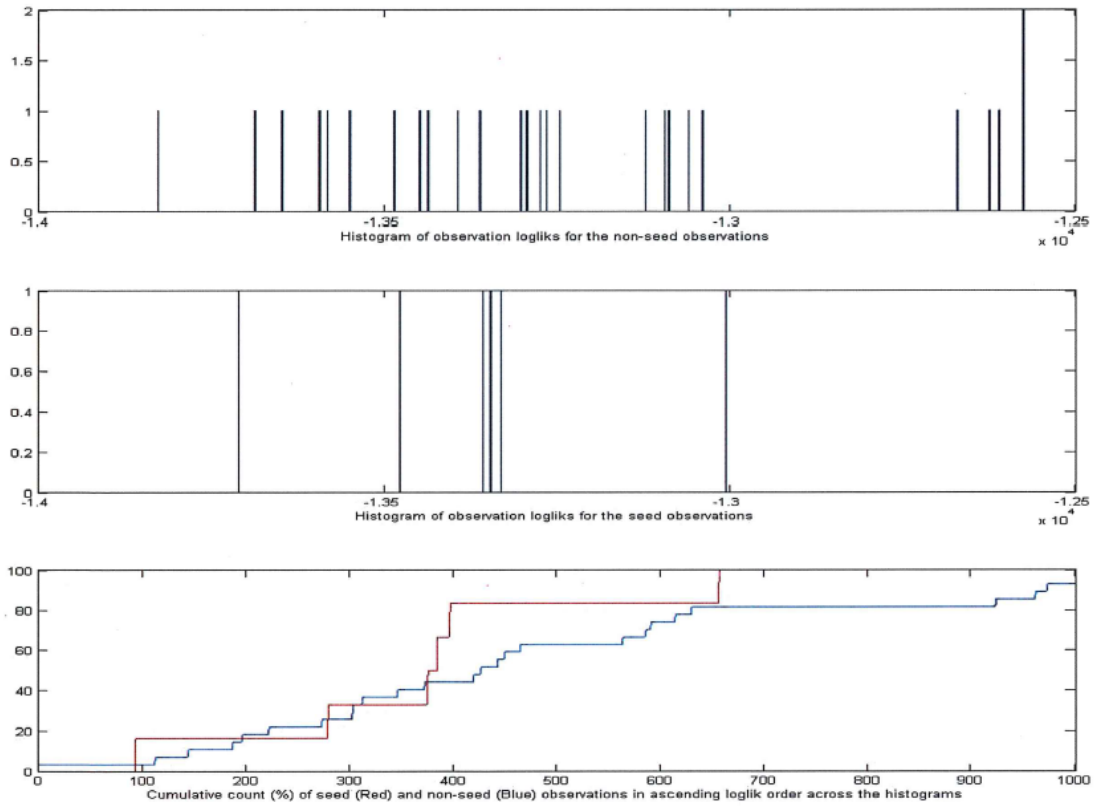

Figure 18 - Distributions of Likelihoods in the HTTP(S) Model

8. From Modelling to Capability

Statistical learning is likely to become of increasing import to network analysts as a means of defeating or moderating the impacts of 'big data', ubiquitous encryption and smart, agile cyber adversaries. Techniques which are unsupervised, behavioural-based, and yet adaptable to specific problem scenarios will be the foundation of future toolsets and have applicability in characterisation, detection and filtering. Hence the continued development of tools such as DMM modelling is important.

In a future deployment, potential applications could include both offline and online (streaming, or real-time) testing. Computationally intensive tasks such as model building (and periodic model refresh) would suit offline processing, but the measurement of likelihood against an existing model is simpler and suitable for stream processing.

Inevitably, maturation of behavioural modelling toward practical use with low false alarm rate will require co-operation between researchers and analysts. For example, in-situ development at an operations centre, with ready access to both current, relevant data and the experience of analysts, could allow fine tuning of features with respect to detection or enrichment of particular behaviours.

9. Conclusions

This work has been sufficient to demonstrate that DMM modelling of network traffic metadata has the potential to assist with problems in source characterisation, cyber-security and volume management. Specific items of merit include:

- Relative ease of convergence, with manageable source counts.
- Behavioural distinctions between observation clusters are often possible and can reveal atypical behaviours within a class.
- Modelling was effective at different scales, revealing behavioural differences when applied to both aggregate backbone traffic and domain-specific traffic such as HTTP(S)-only.
- Modelling could be tuned to a domain of interest (such as the seed-only traffic) in order to filter data to achieve domain enrichment.
- Measures of likelihood could be mined for additional behavioural trends or filtering decisions.
- The approach is unsupervised and suits discovery without signatures. Nonetheless it is likely that model building and analysis would require tuning on a case-by-case basis for each access and problem scenario.

Stronger claims to utility are as yet tempered by limited coverage of the scope of research identified in Section 2. In particular:

- The data management activities of labelling, seeding and filtering had limited treatment here, due largely to an absence of ground truth knowledge and uncertain distinction of the seed behaviours.
- An exhaustive survey of features, and a systematic way of linking the problem statement to features of interest, was lacking.

Future work should follow the general scope of Section 2. Within this, the best resolution of behaviours should result when profiling edge-based observations over successive short windows, which also provides richer structure in the likelihood measurements. To constrain the observation count and ease model building, profiling could be limited to specific domains, with the added benefit of reduced overall behaviour count. Identification of specific problem statements within this remit could enable purposeful experimentation with priors and better feature-to-problem mapping.

Australia's SIGINT mission will confront more data, more encryption, and an ever-changing baseline of 'normal communication'. Managing scale, learning from behaviours, and recognising the truly anomalous will require advanced unsupervised techniques like DMM modelling.

10. References

1. Wikipedia, *Network Management,* https://en.wikipedia.org/wiki/Networkmanagement, 2016.
2. NST Org., *Network Security Toolkit,* http://www.networksecuritytoolkit.org/nst/index.html, 2016.
3. Hastie, T. et.al., *The Elements of Statistical Learning (Data Mining, Inference and Prediction),* Springer, 2008.
4. Ghosh, J., et.al., *An Introduction to Bayesian Analysis - Theory and Methods,* Springer, 2006.
5. Darktrace Inc., *Machine Learning Technology,* https://www.darktrace.com/technology/#machine-learning, 2015.
6. *Johnson, M.,* Bayesian Inference for Dirichlet-Multinomials and Dirichlet Processes, Macquarie University Machine Learning Summer School.
7. Dostalek, L., Kabelova, A., *Understanding TCP/IP,* Packt Publishing, 2006.
8. Cisco Netflow Products, www.cisco.com.
9. RFC 5102, *Information Model for IP Flow Information Export,* 2008.
10. Wikipedia, *Tier* 2 *network,* https://en.wikipedia.org/wiki/Tier2network, 2016.
11. Kim, H. et.al., *Internet Traffic Classification Demystified: Myths, Caveats and the BestPractices,* ACM CoNEXT Conference, 2008.
12. Mathworks Inc., *Matlab,* http://au.mathworks.com/products/matlab/, 2016.
13. Heckerman, D., *A Tutorial on Learning with Bayesian Networks,* MS-TR-95-06, Microsoft Research, March 1995.
14. Norelli, A., *Advanced Persistent Threat Terminology,* DSTO-TN-1355, DSTO, Sept 2014.
15. Norelli, A., *Flow Data Record Analysis for Cyber Discovery: A Survey,* DSTO-GD-0865, DSTO, Jan 2015.
16. SANS Institute Training, *FOR572: Advanced Network Forensics and Analysis,* https://www.sans.org/.
17. Xu, K., Zhang, Z., Bhattacharyya, S., *Profiling Internet Backbone Traffic: Behaviour M odels and Applications,* Proc. of ACM SIGCOMM, 2005.
18. *Franzen, J.,* Bayesian Cluster Analysis - Some Extensions to Non-standard Situations, Stockholm University, 2008.
19. *Cisco Talos,* Cognitive Research: Learning Detectors of Malicious Network Traffic, http://blogs.cisco.com/security/talos/machine-learning-detectors, 2016.

Appendix A Gibbs MCMC DMM Pseudocode

The observations to be modelled are *Xhist,* an N by K array

Specify the estimated source count

S = 32

Initialise the priors

alp = 1xK vector of 1's
bet = 1xS vector of 1's

Initialise the hidden indicator variable Z such that each observation is randomly allocated to one and only one source at a time

Z = NxS array of false
for every row n in N
    Z(n,random) = true
end

Initialise Theta and Phi, sampled from Dirichlet distributions

for every source s in S
    thet(s) ~Dir(alp)
end
phi ~Dir(bet + sum(Z))

Initialise parameter storage for the MCMC iterations

iter = 0
track_phi = empty
track_theta = empty
track_Z = empty

While iterating, count iter and...

Sample the current state of Phi based on the last state of Z

phi ~Dir(bet + sum(Z))

Sample the current state of Theta based on the subset of observations per source (from the last state of Z)

for every source s in S
    let ix be every row where the source is s according to Z
    thet(s) ~Dir(alp + sum(Xhist(ix,:)))
end

From the new states of Phi and Theta, re-estimate the source allocations Z

PZi = NxS array of 0's
for every row n in N
for every source s in S
$PZi(n,s) \propto phi(s) * \prod thet(s)^{Xhist(n,:)}$
end
let the non-zero column in Z(n,:) be that with the largest PZi(n,:)
end

Store the latest iteration of parameters

track_phi{iter} = phi
track_theta{iter} =thet
track_Z{iter} = Z

Produce MAP estimates

map_phi = lxS vector of 0's
for every source s in S
find the histogram of samples of phi(s) from track_phi
let map_ phi(s) be the mode of the histogram
end

map_theta = SxK array of 0's
theta_sk = 1 x iter vector of 0's
for every source s in S
for every feature k in K
find the histogram of samples of theta(s,k) from track_theta
let map_theta(s,k) be the mode of the histogram
end
end

PZi = NxS array of 0's
map Z = NxS array of false
for every row n in N
for every source s in S
$PZi(n,s) \propto map\_phi(s) * \prod map\_thet(s)^{Xhist(n,:)}$
end
let the non-zero column in mapZ(n,:) be that with the largest PZi(n,:)
end

DEFENCE SCIENCE AND TECHNOLOGY GROUP DOCUMENT CONTROL DATA	1. DLM/CAVEAT (OF DOCUMENT)

2. TITLE	3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED LIMITED RELEASE USE (U/L) NEXT TO DOCUMENT CLASSIFICATION)
Bayesian Modelling of Network Traffic Metadata using Dirichlet Multinomial Mixtures	Document (U) Title (U) Abstract (U)
4. AUTHOR(S)	5. CORPORATE AUTHOR
Kevin Harman	Defence Science and Technology Group PO Box 1500 Edinburgh South Australia 5111 Australia

6a. DST GROUP NUMBER	6b. TYPE OF REPORT	7. DOCUMENT DATE
DST-Group-TR-3538	Technical Report	October 2018
8.TASK NUMBER	9.TASK SPONSOR	10. RESEARCH DIVISION
N/A	N/A	Cyber and Electronic Warfare Division

11. MSTC	12. STC
Cyber Sensing and Shaping	Communication Networks Research

13. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release.

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

14. DELIBERATE ANNOUNCEMENT

No limitations

15. CITATION IN OTHER DOCUMENTS

Yes

16. RESEARCH LIBRARY THESAURUS

Network Traffic Profiling, Statistical Learning, Bayesian, Dirichlet

17. ABSTRACT

Statistical theory commends probabilistic modelling techniques for the discovery of latent structure in large datasets not amenable to analysis by inspection. Netflow metadata, for example, may contain latent structure representing different traffic behaviours. The utility of a class of Bayesian models known as Dirichlet multinomial mixtures in discovering such behaviours, and how they might be applied to network analysis problems such as source characterisation, event detection or filtering, is considered herein. Encouragingly, under the right conditions, these models are found to detect and quantify meaningful behavioural distinctions. For an analyst using metadata to mitigate privacy, volume or encryption constraints, but faced with the unpredictable behaviours of cyber adversaries with ever-evolving tools, techniques and procedures, unsupervised learning like Dirichlet mixture modelling could prove a valuable tool.