# Topic modelling of authentication events in an enterprise computer network

Nick Heard
Department of Mathematics, Imperial College London
and Heilbronn Institute for Mathematical Research,
University of Bristol
Email: n.heard@imperial.ac.uk

Konstantina Palla
Department of Statistics
University of Oxford
Email: palla@stats.ox.ac.uk

Maria Skoularidou
Department of Statistics
Athens University of Economics
and Business
Email: m.skoularidou@gmail.com

*Abstract*—The possibility for theft or misuse of legitimate user credentials is a potential cyber-security weakness in any enterprise computer network which is almost impossible to eradicate. However, by monitoring the network traffic patterns, it can be possible to detect misuse of credentials. This article presents an initial investigation into deconvolving the mixture behaviour of several individuals within a network, to see if individual users can be identified. Towards that, a technique used for document classification is deployed, the Latent Dirichlet allocation model. A pilot study is conducted on authentication events taken from real data from the enterprise network of Los Alamos National Laboratory.

## I. Introduction

Statistical anomaly detection methods within cyber security are growing in perceived importance [?]. Data summarising the computer network traffic of an enterprise can be cheaply and routinely gathered, allowing analysts to build statistical models of the normal patterns for data packets passing between users and internet protocol (IP) addresses. Intruders to the network and insider threats can then be potentially detected when their behaviour deviates from that norm. However, the traffic from a computer or IP address is a complex mixture of both automated and human activity, which in turn might represent a mixture of several individuals' activity. Building robust statistical models is therefore a challenging task.

The aim of this article is to use Latent Dirichlet allocation analysis [?], also known as *topic modelling*, of computer network connection traffic data to determine the number of users present. Particular attention is paid here to computer authentication events, which record users authenticating their network credentials on different hosts throughout a working day. A brief pilot study will investigate the feasibility of inferring which users are present when such data are aggregated.

## II. Authentication data from Los Alamos National Laboratory

In 2015, [?] published a comprehensive cyber security data set taken from the computer network of Los Alamos National Laboratory. The full data comprise records of network flows, DNS look ups, user processes and authentication events. A description of the data and links to download are given at http://csr.lanl.gov/data/cyber1. This paper chooses to focus on user authentications on destination computers, for which in total there are 336,806,387 observed events generated by 12,093 unique users authenticating their user credentials on 15,417 unique computers over 58 days. An example record the data is as follows:

```
1 ,C625$@DOM1, U147 @DOM1,C625, C625 ,Negotiate,...
```

The relevant fields are boxed and respectively represent time, user name and destination computer.

Considering authentication connections between users and computers as a bipartite graph, Figs. 1 and 2 show the degree distributions of the users and computers in the data. There are two points to note from these plots: Firstly, the degrees of users are much smaller than computers; users authenticate on just one computer more commonly than any other number, although the median degree is still 24. Secondly, the computers have an approximate power-law with some very high degree computers, the highest being 10,365 which corresponds to over 85% of the authenticating users on the network; the presence of connectivity to such high degree computers is likely to be less informative in identifying the presence of specific users in an aggregated data set, compared to connections to the 1487 computers which have degree 1.
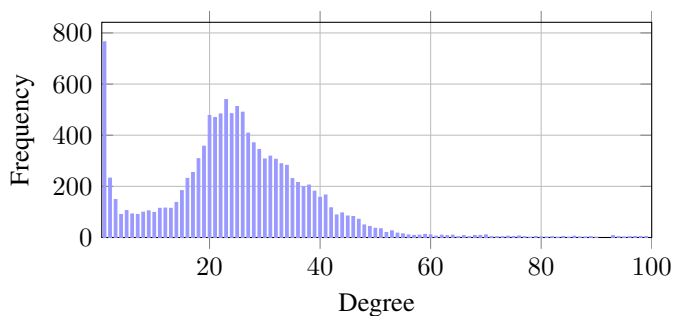


Fig. 1: Degree distribution for LANL users, measured by the number of unique computers on which they authenticated.

Fig. 3 shows the distribution of times of day and the days of week of the authentication event time data. Clear diurnal and weekend effects characteristic of human traffic are present in the data, but it is also notable that there is considerable activity through the nights and across the weekend, suggesting a strong additional presence of automated traffic.
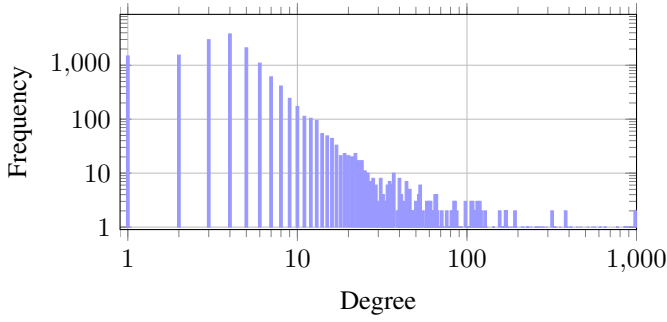
Fig. 2: Log-log plot of degree distribution for LANL computers, measured by the number of unique users authenticating. An approximate power-law relationship can be seen.
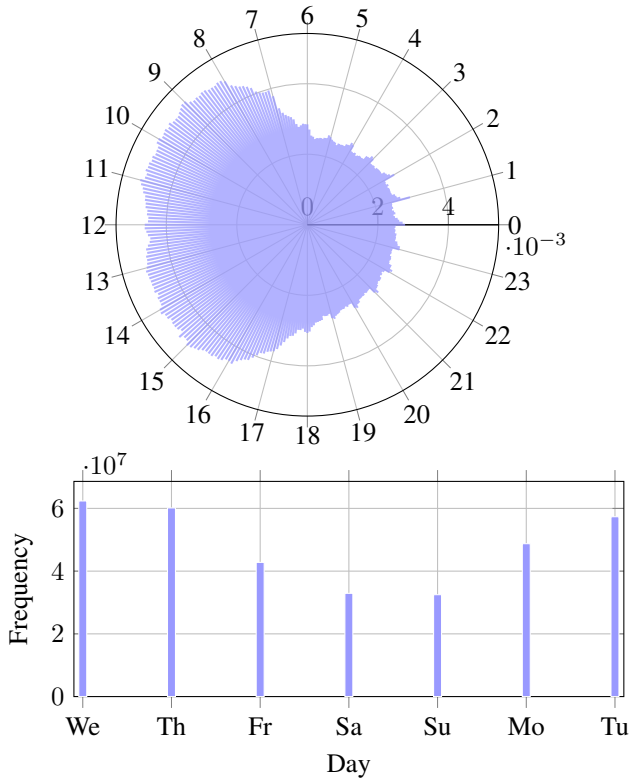


Fig. 3: Distributions of authentication event times. Top: Time of day, in five minute bins. Bottom: Day of week.

## III. Latent Dirichlet allocation modelling

Suppose user connectivity data have been collected across $K$ users over $D$ days. Only a subset of the $K$ users will be active on any given day. On day $d \in \{1, \ldots, D\}$, let $U_d \subseteq U = \{1, \ldots, K\}$ be the unobserved set of users making connections, and let $N_d$ be the total number of connections made that day by those users. The inference task is to learn which users (up to a relabelling) were present each day using only the aggregated connectivity data.

### A. LDA model
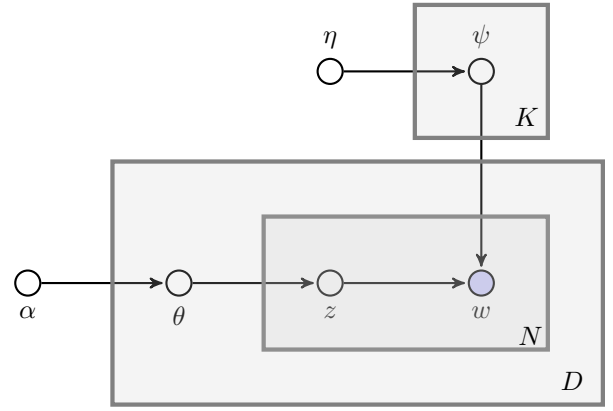
Following the notation of [?] we define



Fig. 4: Graphical representation of the LDA model using plate notation to indicate repetition. The lower outer plate represents documents, while the inner plate represents the repeated choice of users and words within a document.

- A word as the basic unit of discrete data, defined to be an item from a vocabulary set $\mathcal{C}$. A word in this context will be one that indicates 'destination computer', e.g. C625 in the example of Section I.
- A document is a collection of words. In the present context, a document will represent a day of authentication records. Each document $d$ is a sequence of $N_d$ words denoted by $\boldsymbol{w}_d = (w_{d1}, w_{d2}, \ldots, w_{dN_d}) \in \mathcal{C}^{N_d}$.
- A corpus is a collection of $D$ documents denoted by $\mathcal{D} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_D\}$. In this context, the corpus will be the set of records obtained after $D$ days.

The generative mechanism for authentication data from the LDA model, for each document $\boldsymbol{w}_d$ in a corpus $\mathcal{D}$, is as follows:

1) Choose $N_d \sim \text{Poisson}(\xi)$.
2) Choose $\theta_d \sim \text{Dirichlet}(\alpha)$.
3) For each of the $N_d$ words $w_{dn}$:
   - Choose a user from $U$, $z_{dn} \sim \text{Multinomial}(\theta_d)$.
   - Choose a word from $\mathcal{C}$, $w_{dn} \sim \text{Multinomial}(\psi_{z_{dn}})$.

Here $\psi$ is a $K \times |\mathcal{C}|$ matrix where $|\mathcal{C}|$ is the dimension of the vocabulary. The $k$th row of $\psi$, denoted $\psi_k$, is the word distribution of user $k$ and is drawn from $\text{Multinomial}(\eta)$.

So the desired analogue with topic modelling is as follows: Each user is a topic of the LDA model and the vector $\theta_d$ represents the mixing proportions of the users in the daily record $d$. In other words, $\theta_d$ is a vector of length $K$ and each element $\theta_{dk} \in [0, 1]$, for $k \in U$ indicates how active user $k$ is on day $d$, with higher values indicating higher activity. Fig. 4 gives a diagrammatic representation of the full LDA model.

## IV. Application

### A. Subsampled data

Three users were selected from those displaying human-like behaviour, here characterised by zero authentication events occurring over the weekends. For each user a fortnight of data was observed, although besides the weekend effects there were

other days where these users showed no activity, meaning the subsample actually yielded 27 user-days of data; these will serve as the documents for LDA. Note that in this case, each document corresponds to one user, and so the aim is to infer topic distributions $\theta_d$ for each document $d$ such that most of the mass resides in just one component for each $d = 1, \ldots, 27$. A heat map of the sampled data is shown in Fig. 5. Together the three users authenticated on a total of 42 different computers, and so the size of the vocabulary $|\mathcal{C}| = 42$.



Fig. 6: LDA model projections of the user authentication profiles. Each axis represents one of the fitted LDA topics.
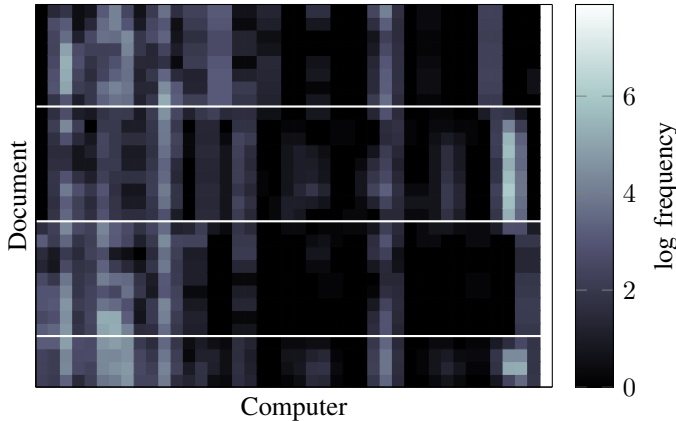


Fig. 5: Heat map of connections to each computer for each authentication event document, with documents sorted into blocks corresponding to the same user. The bottom segment is a mixture of data from two of the users in 4 subsequent days.

### B. Results

A three-topic ($K = 3$) latent Dirichlet allocation model from Section III was fit to the sample data using the python package `gensim`. This package uses a stochastic optimisation variational Bayes approach to estimate the LDA parameters; see [?] for further details. The estimated values of the topic probability distributions $\theta_d$ for each of the 27 documents are shown in Fig. 6. The documents for the three users cluster into three distinct groups of nine documents, located at the vertices of the unit cube; LDA has successfully detected the individual users as the latent topics in the data, without prior knowledge of this information. Note that there is very little variability for U7422 compared to the other users, demonstrating that some users are more easily identifiable than others.

To briefly examine the ability of the fitted model to discriminate mixtures of user data, the next Monday-Friday of data for the users U7442 and U3104 were aggregated. This yielded another four days of data (see the bottom segment of Fig. 5) since both users were not present on the Friday. LDA detected these new data to be mixtures of the two correct users, with negligible mass assigned to U2255. The estimated topic probability distributions for these documents, indicated by the asterisk points in Fig 6, lie in an approximately straight line between the two user clusters. Notably one of the mixture points lies directly on top of the tight cluster for user U7422;
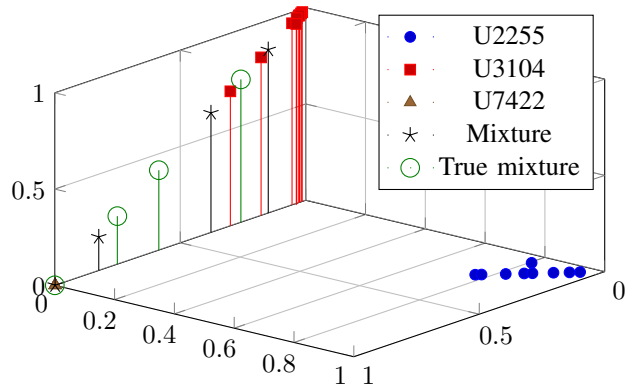
but this is correct, as it transpires that U3104 was also absent on this particular day (Thursday). The true mixing proportions, given by the ratio of events attributable to the two users, are also indicated in Fig 6; these are less well estimated.

### V. CONCLUSION

A first exploratory analysis into using latent Dirichlet allocation models for classifying computer network traffic has been presented. With only a reasonable number of days of training data, the method was able to detect individual users as the underlying *topics* from a topic-modelling perspective.

Further investigations are planned: Firstly to extend the analysis to a larger number of users. This presents no added methodological difficulty, but some initial experimentation has shown the user-topics become more difficult to detect as the number of users increases. It is hypothesised that increasing the number of users will increase the number of days of data required to correctly extract the user-topics. Secondly, the correct number of topics has been assumed known; assuming more topics than needed tends to do little harm (assuming $K = 4$ users here, for example, led to identical inference), but underestimating the number of topics is more problematic.

A later goal, beyond the scope of this article, is anomaly detection. Once users can be identified as topics in a data stream and user presence inferred, more accurate predictive probabilities can be estimated for the next authentication event in the stream. In contrast, without decomposing the mixture of traffic into putative user-topics, the predictive distributions will have higher entropy and strong detection of anomalies is less feasible.

### REFERENCES

[1] N. Adams and N. Heard, *Dynamic Networks and Cyber-Security*, ser. Security Science and Technology. Singapore: World Scientific, 2016, vol. 1.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[3] A. D. Kent, "Comprehensive, Multi-Source Cyber-Security Events," Los Alamos National Laboratory, 2015.

[4] M. Hoffman, D. M. Blei, and F. Bach, "Online learning for latent dirichlet allocation," *Advances in Neural Information Processing Systems*, vol. 23, pp. 856–864, 2010.