New Edge Activity and Anomaly Detection in Computer Networks

Silvia Metelli [1,2] Nick Heard [1,3]

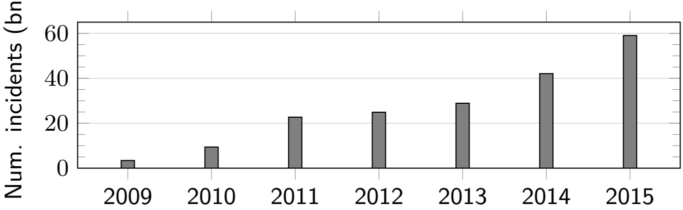
[1] Imperial College London
[2] The Alan Turing Institute
[3] Heilbronn Institute for Mathematical Research

2nd IMA and OR Society Conference on Mathematics of Operational Research, 25-26 April 2019 - Birmingham

Overview

Motivation: Increasingly sophisticated, multi-stage cyber-attacks
e.g. *WannaCry 2017: 230,000 computers in 150 countries*

Global number of cyber incidents 2009-2015 (source: PwC)



(UK, 2018: 7407 breaches in businesses → GOV UK invested ∼1.9b)

Goals: Monitor the computer network by modelling new edge formation at scale & identifying latent network structure

Challenges: Computational speed and scalability

Intrusion Detection Approach

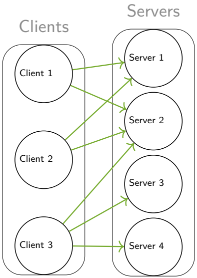
Anomaly-based: deviation from a model of the normal state
(can detect new attacks in contrast to signature-based methods, Patcha et al., 2007)

$\Downarrow$ need reliable, scalable models of *normal* state

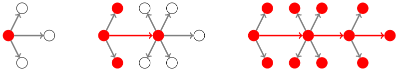
Modelling approach to the evolution of new edges in the computer network:

New Edges: connections between a client and server pair not previously observed

Given a bipartite graph $G(t)$ and
a history $H(t)$ of all connections at time t
$\downarrow$
interest in P(**new edges** at $t+1 |\ H(t)$)



Modelling New Edges



Step 1 Step 2 Step 3

New edges can be

rarely, signal of anomaly

regularly, formed by uninfected hosts

- ▷ we need to understand the rate of occurrence of new edges
- ▷ we need to predict the **identity** of new edges (also identifying **latent structure**)

Why latent structure?

Do similar clients connect to similar servers? If so, it can be predictive of similar future interactions ⇒ build a model for new edges which also considers latent structure

New Edge Intensity

We propose a model for the conditional intensity of observing a new edge (x, y), $x \in X$, $y \in Y$:

$$\lambda_{xy}(t) = r(t) \exp\{\alpha \cdot (N_x^+(t), N_y^-(t), I_{x,1}(t), I_{x,2}(t)) + \beta_{xy} \cdot Z_{xy}(t)\} \times \mathbb{1}_{(X \times Y) \setminus G_t}\{(x, y)\}$$

- $r(t)$: 'seasonal' baseline hazard
- $N_x^+(t), N_y^-(t)$: time-varying in-degree of x and out-degree of y
- $I_{x,1}(t), I_{x,2}(t)$: time-varying indicators of new edge 'burstiness'
- $Z_{xy}(t)$: **matrix of attraction** $x \leftrightarrow y$ (similarity between clients and servers)

two formulations of $Z_{xy}(t)$

Cluster membership indicators

Dot-products of latent feature positions

1. Cluster Formulation (hard-thresholding)

Simultaneous biclustering of clients and servers:

$\mathbb{C} = \{C_1, \ldots, C_L\}$ partition of the client set X
$\mathbb{S} = \{S_1, \ldots, S_M\}$ partition of the server set Y

$$\boxed{Z_{xy}(t) = \left(N^{+}_{x|\mathbb{S}(y)}(t), N^{-}_{y|\mathbb{C}(x)}(t)\right)}$$

outdegree of y restricted to cluster $l(x) \in \mathbb{C} \rightarrow N^{+}_{x|S}(t) = \sum_{n \geq 1} \mathbb{1}_{[0,t)}(t'_n)\mathbb{1}_{x}(x'_n)\mathbb{1}_{S}(y'_n)$

indegree of x restricted to $m(y) \in \mathbb{S} \rightarrow N^{-}_{y|C}(t) = \sum_{n \geq 1} \mathbb{1}_{[0,t)}(t'_n)\mathbb{1}_{x}(y'_n)\mathbb{1}_{C}(x'_n)$

limitations: single, finite representation + each data point only to one cluster

flexible embedding: potentially infinite number of latent features

↓

Indian Buffet Process (IBP)

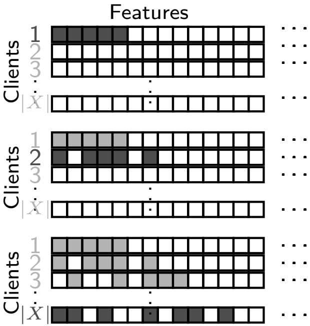
$K =$ number of latent features

$U = (u_1, \dots, u_{|X|}) \in \mathbb{R}^{|X| \times K}$ (clients)

$V = (v_1, \dots, v_{|Y|}) \in \mathbb{R}^{|Y| \times K}$ (servers)

$$\boxed{Z_{xy}(t) = u_x \cdot {v_y}^T}$$

automatically accounts for biclusters



Two-step Inference

Bayesian framework $\rightarrow$ posterior inference with MCMC

MCMC depends on starting values $\rightarrow$ need 'good' initial latent structure:

Surrogate Model for Cluster form.: Model-based Agglomerative Biclustering
Surrogate Model for Latent form.: Sparse SVD + stability selection

Updating scheme

$\mathbf{1^{st}}$ step: initial latent structure via surrogate model

$\mathbf{2^{nd}}$ step: jointly update initial structure and model parameters through MCMC

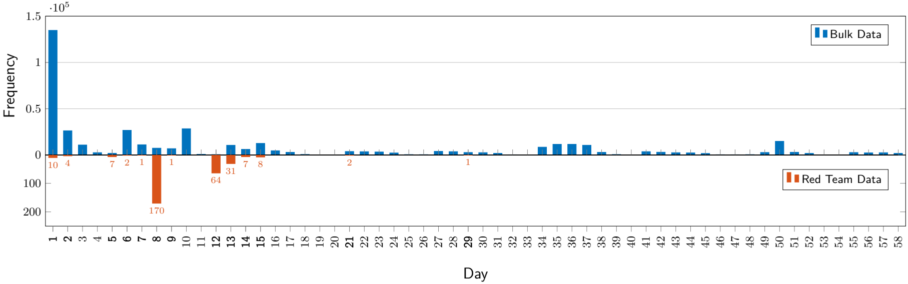
Cyber-security application

Bulk:

- 1,648,275,307 events in total (58 days of traffic)
- 16,230 clients – 15,417 servers

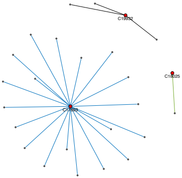
Red Team:

- penetration testing: subset labelled as known compromised events
- 48,079 of the total records: 4 compromised clients

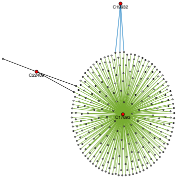


Red Team

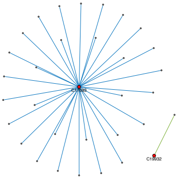
Client computer	Frequency		Unique Server computers	
	Red Team	Total	Red Team	Total
C17693	701	1717	**296**	534
C18025	3	101	**1**	29
C19932	19	10,008	**8**	30
C22409	26	36,253	**3**	31



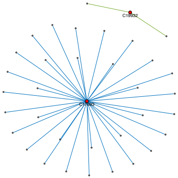
Week 1



Week 2



Week 3



Week 4

Model Prediction Performance

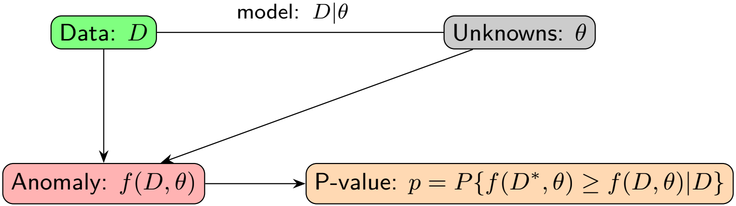
Method tested under both cluster and latent formulation:

- 15 sample repetitions (all events from 1000 randomly sampled clients)
- positive model coefficients: strong impact of degree and latent structure
- out-of-training log likelihood on the last 10,000 events

Model	**Log Likelihood**	**Iteration Time**
Cluster	-18804.34	81.4s
Latent-feature (IBP)	-18379.93	131.7s

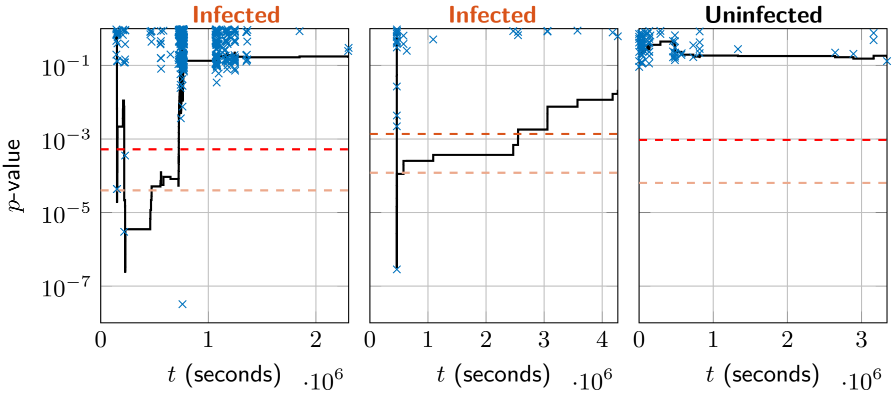
We find that the latent feature model outperforms the cluster model

Anomaly-detection



H$_0$: normal behaviour
H$_1$: departure from model of normality learned

Anomaly-detection

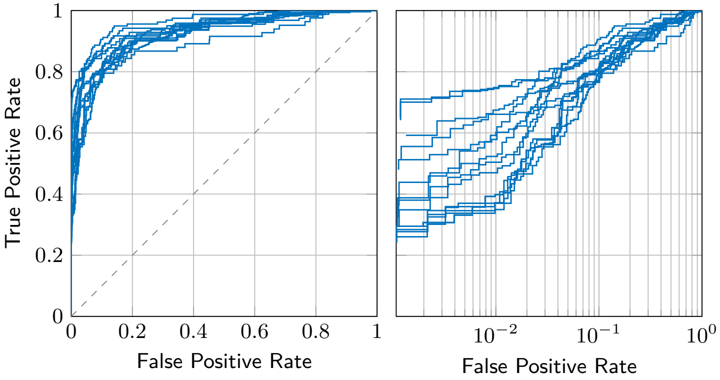


$$p_n = \frac{\sum_{(x,y)\notin G_{t'_n}} \lambda_{xy}(t'_n)\mathbb{1}_{(0,\lambda_{x'_n y'_n}(t'_n)]}\{\lambda_{xy}(t'_n)\}}{\sum_{(x,y)\notin G_{t'_n}} \lambda_{xy}(t'_n)}$$

$$s_x(t) = \bar{\chi}^2_{2\{1+N_x^+(t)\}}\left(-2\sum_{n\geq 1}\mathbb{1}_{[0,t)}(t'^x_n)\log p^x_n\right) \text{ s.t. } \inf_{t\geq 0}\; s_x(t)$$

Anomaly-detection

ROC curves for each client (for each sample repetition)



Conclusion

We proposed a Bayesian model and anomaly-detection method:

1) modelling rate and identity of new edges and simultaneously
2) detecting latent network structure to aid new edge prediction

Application to computer network data:

- good prediction performance
- latent formulation outperforms cluster formulation
- anomaly-detection with good false/positive rate
- successfully detected two known compromised clients

Further Research:
- adapt the choice of the construction of the control chart
- exploit faster inference methods (e.g. variational inference)

References

Kent, A. D. User-computer authentication associations in time, Los Alamos National Laboratory, http://dx.doi.org/10.11578/1160076, 2014.

Patcha, A. and Park, J. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, 51(12): 3448-3470, 2007.

Cox, D. R. Regression models and life-tables (with discussion), *J. R. Statist. Soc. B*, 34, 187-220, 1972.

Griffiths, T. L. and Ghahramani, Z. Infinite latent feature models and the Indian buffet process. *NIPS*, 2005.

Thank you!