Stochastic Blockmodels as an unsupervised approach to detect botnet infected clusters in networked data

Mark Patrick Roeling & Geoff Nicholls

Department of Statistics
University of Oxford

Data Science for Cyber-Security, September 2017

Definition

Botnets consist of devices connected to the internet, supervised by a botnet owner, performing malicious tasks.

Command and Control (C&C) and Peer to Peer (P2P) types. P2P more difficult to identify given the pattern of communication.

Botnets:

- usually propagate through malware based infection of computers
- attractive and useful tools for criminal purposes
- Distributed Denial of Service (DDoS)
- fraud and theft of data or computational resources

Features and methods

Detection methods: malware analysis or differentiating normal versus malicious traffic with machine learning algorithms

Example presented features (host-based or flow-based):

- average payload packet length
- average bits per second
- ratio between the number of incoming packets over the number of outgoing packets
- connection duration

Methods used: decision trees, distance based clustering, support vector machines, perceptrons, neural networks, bayesian methods, clustering based on local shrinking.

Problems

Most parametric machine learning methods:

- neglect the linked or networked structure of the data
- assume conditional independence of the botnet / non-botnet status given node-based traffic summary statistics

Network data are typically collapsed for every node:
Example: position of a node based on number and type of neighbours.

Validation sets often include the same botnets = evaluation against a model specifically tuned to bots in training data.

Study aim

Aim: extend the application of SBMs to cybersecurity:

- fit SBMs to a capture of network data including botnet infected machines
- SBMs is unsupervised and works on networked data, overcoming problems of earlier studies
- discover a latent class or multiple classes of malicious traffic as a subset of all classes
- these classes consist of clusters including bots based on collective behaviour

Stochastic BlockModels I

- widely used as canonical model for community detection
- extensions of regular latent variable models to networked data
- allow partitioning of vertices (users or addresses on the internet) of a graph into clusters with high connectivity
- cluster membership is inferred from the edge pattern

Botnet dataset consists of IP and DNS addresses (vertices) that exist on the internet. These connections can be represented by a directed binary/digraph graph G consisting of a set of g nodes.

Stochastic BlockModels II

i and j = nodes indices (from 1 up to n)
q and l = denoting class indices (from 1 up to Q)
We consider Q classes on nodes.
For a single relation between two addresses, the adjacency matrix is given by:

$$X = (x_{ij}) \tag{1}$$

with i and j both $\in \{1, ..., g\}$, and

$$x_{ij} = \begin{cases} 1 & \text{if } i \text{ connected to } j, \\ 0 & \text{otherwise.} \end{cases}$$

Data reduced to 1 adjacency matrix, where every ij pair (where $i \neq j$) was 1 if at least one flow occurred between i and j.

Stochastic Blockmodels III

Let the membership matrix Z_{iq} be 1 if i is a member of class q for $i \in \{1, n\}$ and $q \in \{1, Q\}$.

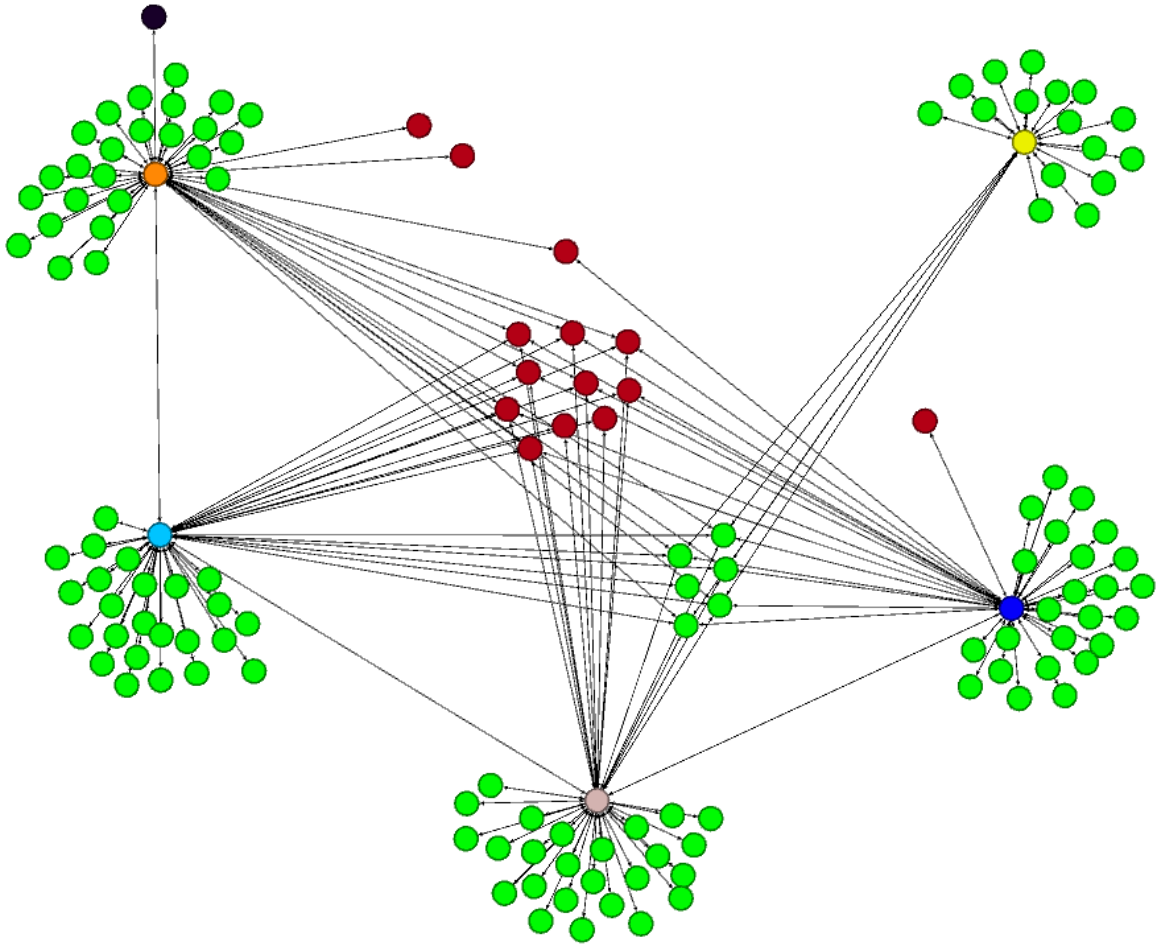
Model is acquired by obtaining the distribution of each edge (i, j) conditional on the membership of node i in the q-th class and node j in the l-th class.

$$X_{ij}|Z_{iq}Z_{jl} = 1 \overset{ind.}{\sim} \mathrm{B}(\pi_{ql}) \tag{2}$$

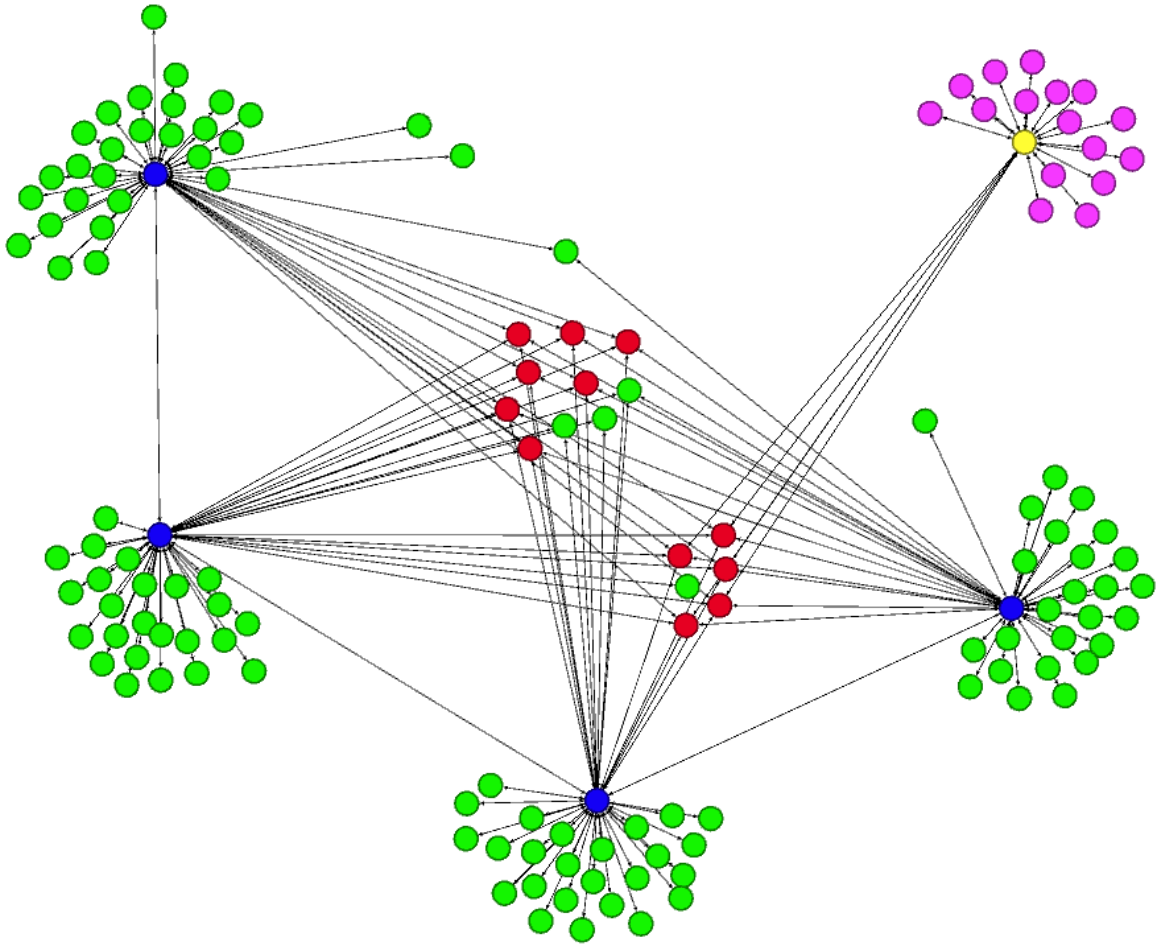
with $(i, j) \in \{1, ..., g\}$ and i never equal to j.

Estimation procedure used variational expectation maximization approach Mariadassou et al. (2010), and Integrated Classification Likelihood from Biernacki et al. (2000).

Simulation study: input network



Simulation study: SBM network Figure



Simulation study: membership Table of SBM

Node Type	class 1	class 2	class 3	class 4	class 5
infected users	0	0	4	0	0
uninfected users	0	1	0	0	0
non-malicious	99	0	0	16	5
malicious	7	0	0	0	7
connector	1	0	0	0	0

Table: Frequencies of the different nodes across classes:
Class 1 = mostly non-malicious nodes
Class 2 = the uninfected user
Class 3 = all four infected users
Class 4 = includes only non-malicious nodes
Class 5 = non-malicious and malicious nodes

Data

Data made available from the University of Victoria: collection of neutral / background data and 4 samples of botnet flows.

The neutral data were collected from the Traffic Lab at Ericsson Research in Hungary and from the Lawrence Berkeley National Lab (LBNL). The Ericsson Lab dataset contains general traffic from: HTTP browsing, gaming streams, and bittorrent packets.

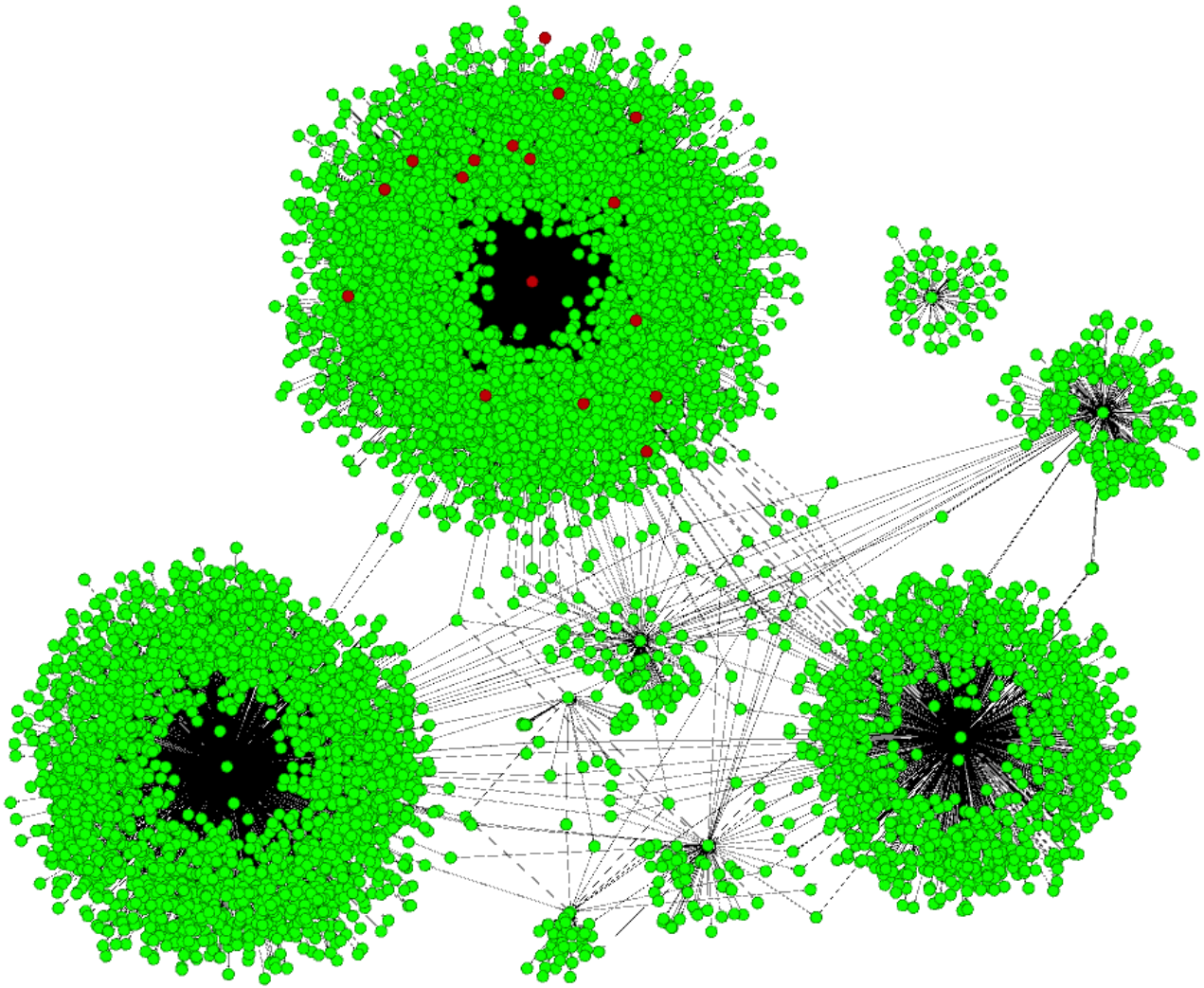
A capture of Zeus botnet traffic was included (C&C and P2P). Biggest and most well known botnets running on Windows, spreads through drive-by-downloads and phishing (3.6m PCs in USA in 2009).

Some Descriptives

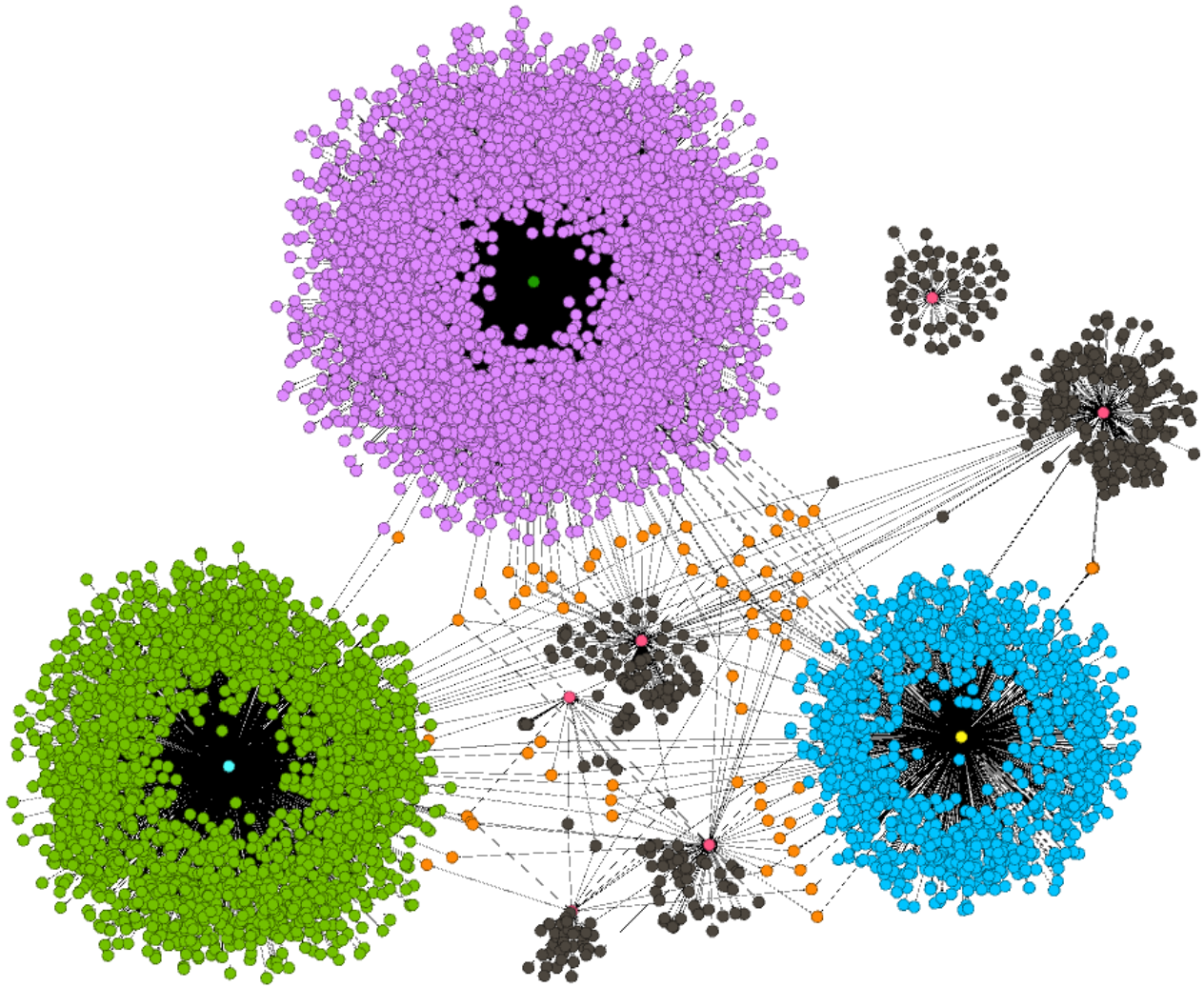
Data were collected in separate environments and the addresses of the botnet data have been mapped to match the addresses of the neutral data so the connections seem to occur within the same network.

1. Unique pairs of connections (13609 in neutral data and 28 botnet data).
2. Neutral 9274 unique nodes, botnet = 18 unique nodes.
3. Merge matrices (some nodes overlapped, eg. IP address 172.16.2.12 was involved in both non-malicious as malicious activity).
4. Transform matrices into an adjacency matrix.

Botnet data: original classification



Botnet data: SBM classification



Simulation study: membership Table

class	#nodes	#neutral	#Zeus	%botnet
1	1144	1144	0	-
2	1	1	0	-
3	5166	5150	16	94.1%
4	6	6	0	-
5	2	1	1	5.88%
6	1	1	0	-
7	2425	2425	0	-
8	81	81	0	-
9	465	465	0	-

Discussion

- SBM successful in Simulation study but not convincing with the Univ. of Victoria dataset
  - Data were collected in separate environments and mapped a posteriori
  - Peer to peer structure largely absent and Zeus capture was small
- One of the downsides of current detection methods in C&C botnets is the requirement of at least 2 infected machines. The simulation study shows that this unsupervised approach also needs 2 machines (to identify shared malicious addresses)

Thank you for your attention.

Email Geoff: nicholls@stats.ox.ac.uk

Email Mark: mark.roeling@stats.ox.ac.uk