Multilayer Block Models for Exploratory Analysis of Computer Event Logs

Corentin Larroche

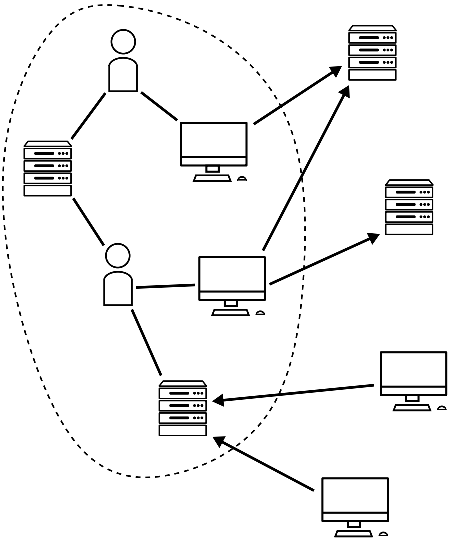
corentin.larroche@ssi.gouv.fr

French National Cybersecurity Agency (ANSSI), Paris, France

Complex Networks '22, Palermo, Italy

November 9th, 2022

Problem definition – Computer network monitoring

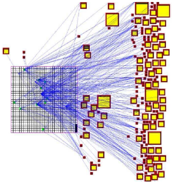


Event logs

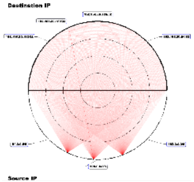
- Record **various types** of activity
- Many events can be seen as **interactions** between entities
- Here, we focus on **authentications** and **network flows**

- Massive amount of data
- Goal: quickly **explore** and **understand** their content, and uncover **suspicious behaviors**

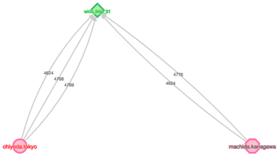
Related work – Visualization tools



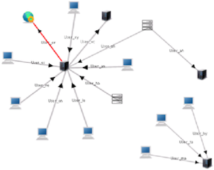
VISUAL [Ball et al., 2004]



FloVis [Taylor et al., 2009]

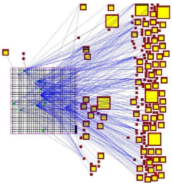


LogonTracer [Tomonaga, 2017]

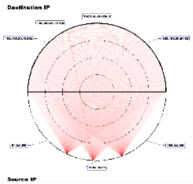


APTHunter [Siadati et al., 2016]

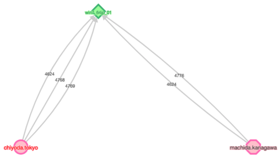
Related work – Visualization tools



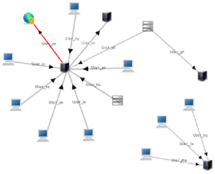
VISUAL [Ball et al., 2004]



FloVis [Taylor et al., 2009]



LogonTracer [Tomonaga, 2017]



APTHunter [Siadati et al., 2016]

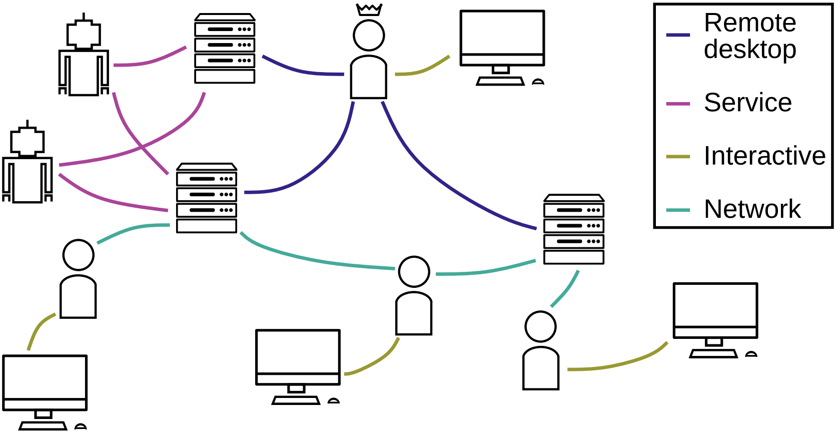
Problem

Displaying everything **does not scale** well!

- Need to **summarize** the graphs

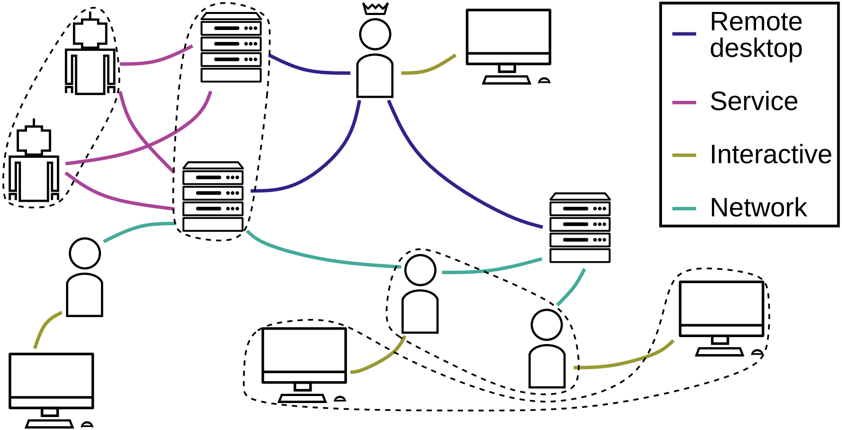
Summarizing the data

Intuition: many nodes have **similar connectivity patterns**.



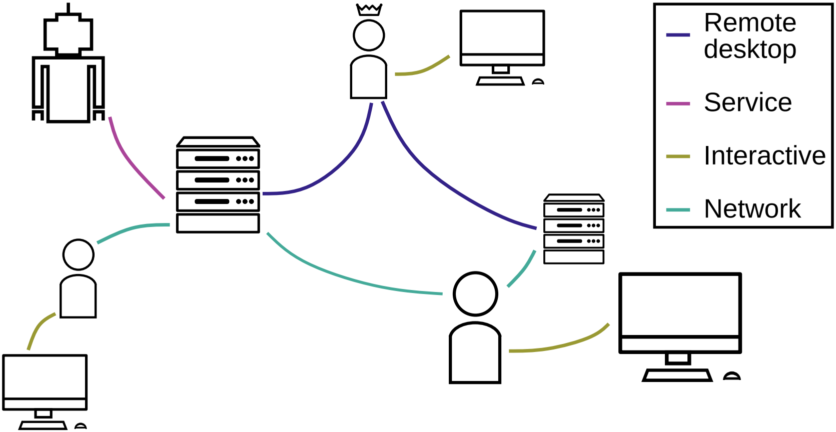
Summarizing the data

Intuition: many nodes have **similar connectivity patterns**.



Summarizing the data

Intuition: many nodes have **similar connectivity patterns**.



Formal definitions and model description

Definitions

Let $\mathcal{U}, \mathcal{V}$ be the **top** and **bottom** node sets, respectively. Assume there are L **edge types**. We consider a **bipartite multiplex graph** $\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subset \mathcal{U} \times \mathcal{V} \times [L]$ is the edge set. For each type $\ell \in [L]$, the biadjacency matrix for layer ℓ is denoted $\mathbf{B}^{(\ell)} = (b_{ij}^{(\ell)})$.

Generative model: multilayer extension of the **Poisson latent block model** [Govaert and Nadif, 2010].

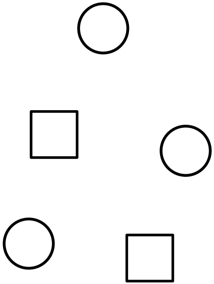
Formal definitions and model description

Definitions

Let $\mathcal{U}, \mathcal{V}$ be the **top** and **bottom** node sets, respectively. Assume there are L **edge types**. We consider a **bipartite multiplex graph** $\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subset \mathcal{U} \times \mathcal{V} \times [L]$ is the edge set. For each type $\ell \in [L]$, the biadjacency matrix for layer ℓ is denoted $\mathbf{B}^{(\ell)} = (b_{ij}^{(\ell)})$.

Generative model: multilayer extension of the **Poisson latent block model** [Govaert and Nadif, 2010].

- H top clusters, K bottom clusters



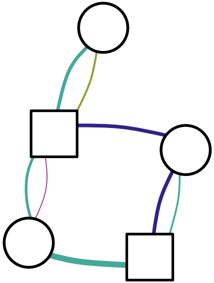
Formal definitions and model description

Definitions

Let $\mathcal{U}, \mathcal{V}$ be the **top** and **bottom** node sets, respectively. Assume there are L **edge types**. We consider a **bipartite multiplex graph** $\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subset \mathcal{U} \times \mathcal{V} \times [L]$ is the edge set. For each type $\ell \in [L]$, the biadjacency matrix for layer ℓ is denoted $\mathbf{B}^{(\ell)} = (b_{ij}^{(\ell)})$.

Generative model: multilayer extension of the **Poisson latent block model** [Govaert and Nadif, 2010].

- H top clusters, K bottom clusters
- For each $\ell \in [L]$, let $\boldsymbol{\Theta}^{(\ell)} = (\theta_{hk}^{(\ell)})$ be a **cluster connectivity matrix**.



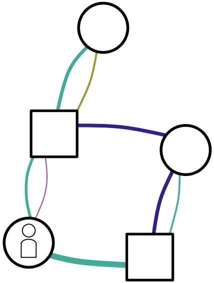
Formal definitions and model description

Definitions

Let $\mathcal{U}, \mathcal{V}$ be the **top** and **bottom** node sets, respectively. Assume there are L **edge types**. We consider a **bipartite multiplex graph** $\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subset \mathcal{U} \times \mathcal{V} \times [L]$ is the edge set. For each type $\ell \in [L]$, the biadjacency matrix for layer ℓ is denoted $\mathbf{B}^{(\ell)} = (b_{ij}^{(\ell)})$.

Generative model: multilayer extension of the **Poisson latent block model** [Govaert and Nadif, 2010].

- H top clusters, K bottom clusters
- For each $\ell \in [L]$, let $\mathbf{\Theta}^{(\ell)} = (\theta_{hk}^{(\ell)})$ be a **cluster connectivity matrix**.
- $\forall i \in \mathcal{U}$, draw cluster $U_i \sim \mathrm{Mult}(\boldsymbol{\pi})$.



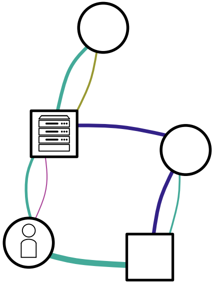
Formal definitions and model description

Definitions

Let $\mathcal{U}, \mathcal{V}$ be the **top** and **bottom** node sets, respectively. Assume there are L **edge types**. We consider a **bipartite multiplex graph** $\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subset \mathcal{U} \times \mathcal{V} \times [L]$ is the edge set. For each type $\ell \in [L]$, the biadjacency matrix for layer ℓ is denoted $\mathbf{B}^{(\ell)} = (b_{ij}^{(\ell)})$.

Generative model: multilayer extension of the **Poisson latent block model** [Govaert and Nadif, 2010].

- H top clusters, K bottom clusters
- For each $\ell \in [L]$, let $\boldsymbol{\Theta}^{(\ell)} = (\theta_{hk}^{(\ell)})$ be a **cluster connectivity matrix**.
- $\forall i \in \mathcal{U}$, draw cluster $U_i \sim \mathrm{Mult}(\boldsymbol{\pi})$.
- $\forall j \in \mathcal{V}$, draw cluster $V_j \sim \mathrm{Mult}(\boldsymbol{\rho})$.



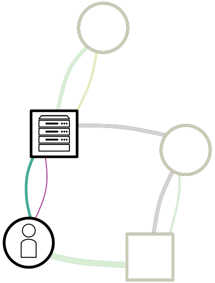
Formal definitions and model description

Definitions

Let $\mathcal{U}, \mathcal{V}$ be the **top** and **bottom** node sets, respectively. Assume there are L **edge types**. We consider a **bipartite multiplex graph** $\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subset \mathcal{U} \times \mathcal{V} \times [L]$ is the edge set. For each type $\ell \in [L]$, the biadjacency matrix for layer ℓ is denoted $\mathbf{B}^{(\ell)} = (b_{ij}^{(\ell)})$.

Generative model: multilayer extension of the **Poisson latent block model** [Govaert and Nadif, 2010].

- H top clusters, K bottom clusters
- For each $\ell \in [L]$, let $\mathbf{\Theta}^{(\ell)} = (\theta_{hk}^{(\ell)})$ be a **cluster connectivity matrix**.
- $\forall i \in \mathcal{U}$, draw cluster $U_i \sim \mathrm{Mult}(\boldsymbol{\pi})$.
- $\forall j \in \mathcal{V}$, draw cluster $V_j \sim \mathrm{Mult}(\boldsymbol{\rho})$.
- $\forall (i, j, \ell) \in \mathcal{U} \times \mathcal{V} \times [L]$, draw edge indicator $b_{ij}^{(\ell)} \sim \mathrm{Poisson}(\mu_i \nu_j \theta_{U_i V_j}^{(\ell)})$.



Model inference and selection

Cluster assignments and model parameters are inferred through **maximum likelihood estimation**.

- Goal: maximize the complete data log-likelihood

$$L_{\mathrm{C}} = \sum_{i} \log \pi_{U_i} + \sum_{j} \log \rho_{V_j} + \sum_{i,j,\ell} \left\{ b_{ij}^{(\ell)} \log \left(\mu_i \nu_j \theta_{U_i V_j}^{(\ell)} \right) - \mu_i \nu_j \theta_{U_i V_j}^{(\ell)} \right\}$$

Model inference and selection

Cluster assignments and model parameters are inferred through **maximum likelihood estimation**.

- Goal: maximize the complete data log-likelihood

$$L_{\mathrm{C}} = \sum_{i} \log \pi_{U_i} + \sum_{j} \log \rho_{V_j} + \sum_{i,j,\ell} \left\{ b_{ij}^{(\ell)} \log \left(\mu_i \nu_j \theta_{U_i V_j}^{(\ell)} \right) - \mu_i \nu_j \theta_{U_i V_j}^{(\ell)} \right\}$$

- We adapt the **variational EM** procedure of [Govaert and Nadif, 2010]:
  (i) Estimate node activities $\boldsymbol{\mu}, \boldsymbol{\nu}$ from the marginal totals of $\mathbf{B}^{(1:L)}$
  (ii) Introduce **soft cluster assignment** matrices $\mathbf{U} \in [0,1]^{|\mathcal{U}| \times H}$ and $\mathbf{V} \in [0,1]^{|\mathcal{V}| \times K}$
  (iii) Alternately optimize $\mathbf{U}$, $\mathbf{V}$ and $\boldsymbol{\Theta}^{(1:L)}$
  (iv) Round $\mathbf{U}$ and $\mathbf{V}$ to obtain hard cluster assignments

Model inference and selection

Cluster assignments and model parameters are inferred through **maximum likelihood estimation**.

- Goal: maximize the complete data log-likelihood

$$L_{\mathrm{C}} = \sum_{i} \log \pi_{U_i} + \sum_{j} \log \rho_{V_j} + \sum_{i,j,\ell} \left\{ b_{ij}^{(\ell)} \log \left(\mu_i \nu_j \theta_{U_i V_j}^{(\ell)} \right) - \mu_i \nu_j \theta_{U_i V_j}^{(\ell)} \right\}$$

- We adapt the **variational EM** procedure of [Govaert and Nadif, 2010]:
  (i) Estimate node activities $\boldsymbol{\mu}, \boldsymbol{\nu}$ from the marginal totals of $\mathbf{B}^{(1:L)}$
  (ii) Introduce **soft cluster assignment** matrices $\mathbf{U} \in [0,1]^{|\mathcal{U}| \times H}$ and $\mathbf{V} \in [0,1]^{|\mathcal{V}| \times K}$
  (iii) Alternately optimize $\mathbf{U}$, $\mathbf{V}$ and $\boldsymbol{\Theta}^{(1:L)}$
  (iv) Round $\mathbf{U}$ and $\mathbf{V}$ to obtain hard cluster assignments

Model selection

The number of clusters is selected through **grid search** by maximizing the **integrated completed likelihood** (ICL [Biernacki et al., 2000]),

$$\mathrm{ICL} \propto 2L_{\mathrm{C}} - (H-1)\log|\mathcal{U}| - (K-1)\log|\mathcal{V}| - LHK\log(L|\mathcal{U}||\mathcal{V}|)$$

Dataset – VAST Challenge 2013 MC3

Two weeks of **simulated network flows** between an enterprise network and external hosts, with **several attacks** (DDoS, port scans, botnet infection, data exfiltration).

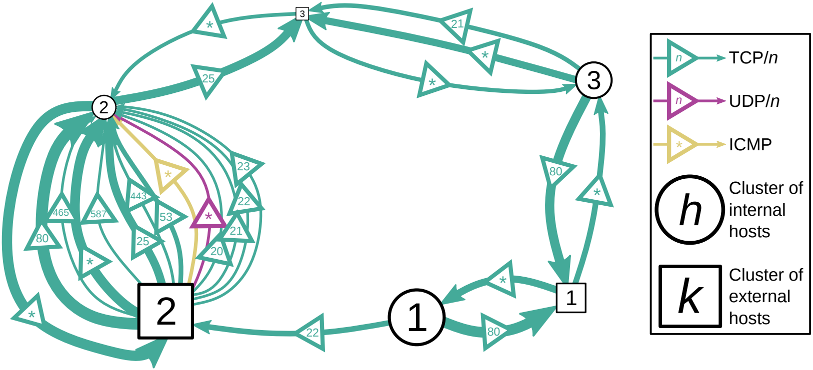
Flow=(@IP$_{\mathrm{src}}$, @IP$_{\mathrm{dst}}$, protocol, Port$_{\mathrm{dst}}$)

Case 1: internal source, external destination

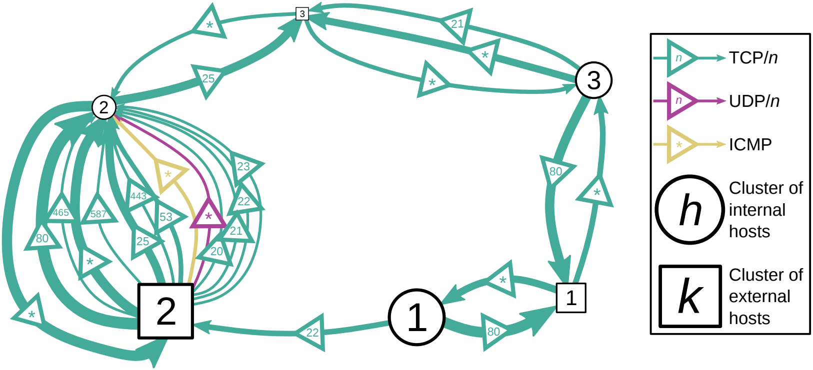
Case 2: external source, internal destination

- 1,220 internal hosts (top nodes)
- 200 external hosts (bottom nodes)
- 18 edge types (dest. port restricted to 10 well-known ports and one "Other port" token)
- 26,597 edges

First case study – Network flows (results)



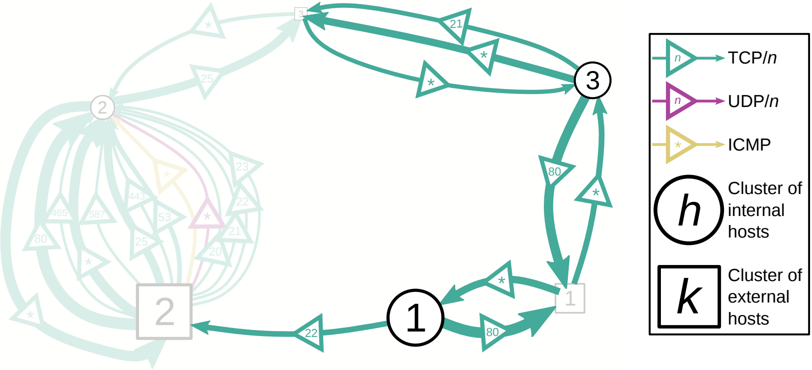
First case study – Network flows (results)



Relevant clusters:

Supicious behaviors:

First case study – Network flows (results)

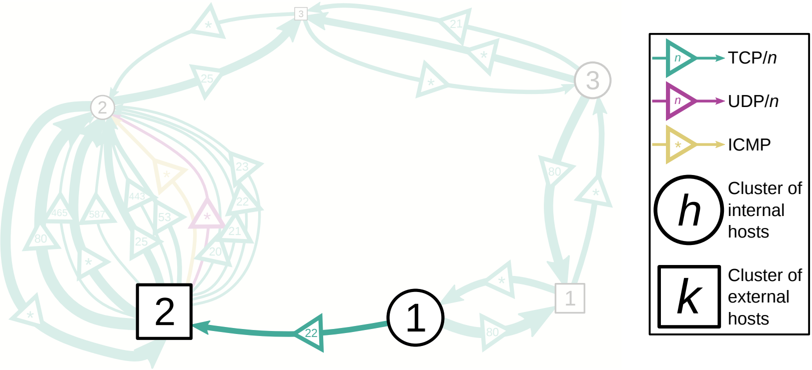


Relevant clusters:

- Internal workstations

Supicious behaviors:

First case study – Network flows (results)



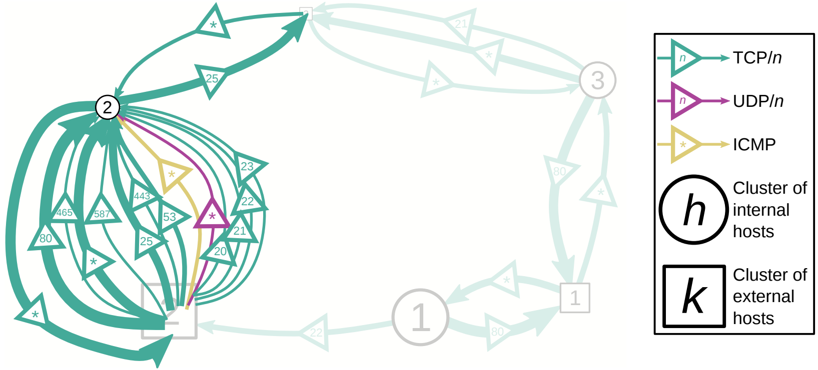
Relevant clusters:

- Internal workstations

Supicious behaviors:

- Outbound SSH traffic from 8 internal hosts to an external host (botnet C&C)

First case study – Network flows (results)



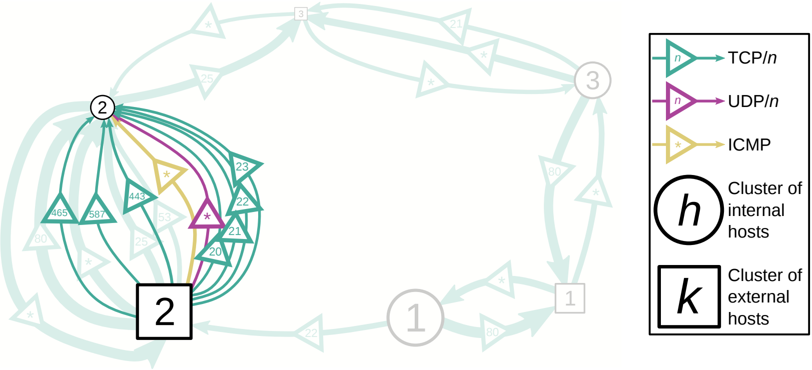
Relevant clusters:

- Internal workstations
- Internal servers

Supicious behaviors:

- Outbound SSH traffic from 8 internal hosts to an external host (botnet C&C)

First case study – Network flows (results)



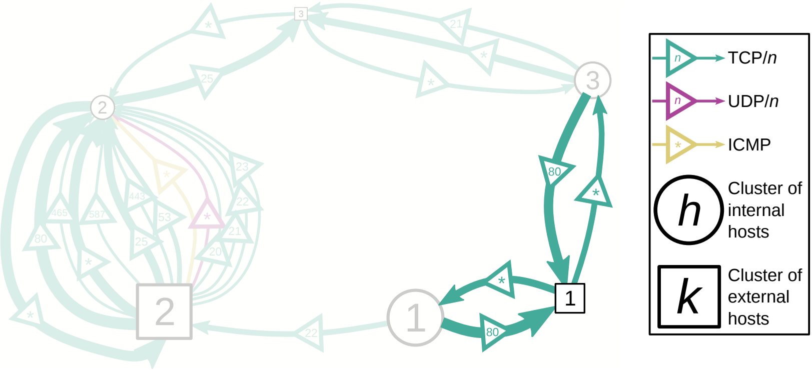
Relevant clusters:

- Internal workstations
- Internal servers

Supicious behaviors:

- Outbound SSH traffic from 8 internal hosts to an external host (botnet C&C)
- Many ports with few connections (port scans)

First case study – Network flows (results)



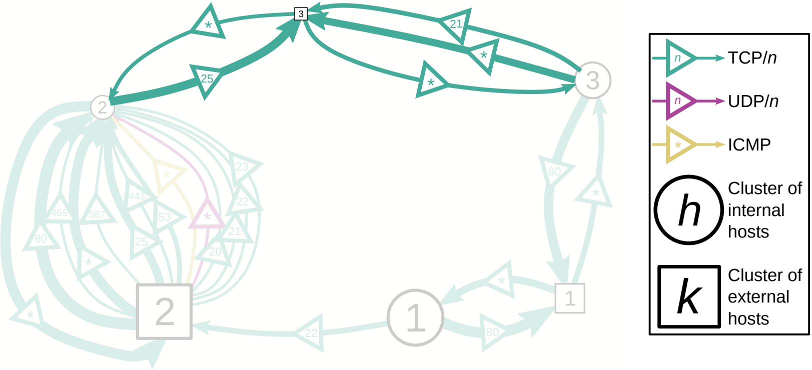
Relevant clusters:

- Internal workstations
- Internal servers
- External Web servers

Supicious behaviors:

- Outbound SSH traffic from 8 internal hosts to an external host (botnet C&C)
- Many ports with few connections (port scans)

First case study – Network flows (results)



Relevant clusters:

- Internal workstations
- Internal servers
- External Web servers
- External FTP and mail servers

Supicious behaviors:

- Outbound SSH traffic from 8 internal hosts to an external host (botnet C&C)
- Many ports with few connections (port scans)

Second case study – Authentication logs (description)

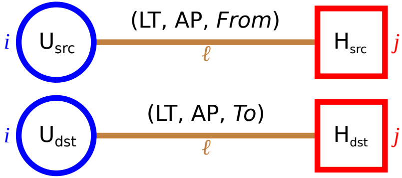
Dataset – "Comprehensive, Multi-Source Cyber-Security Events"

58 days of **authentication logs** from a **real enterprise network**, with labelled events corresponding to a **red team exercise**.

Event=(U_{src}, U_{dst}, H_{src}, H_{dst}, AuthPkg, LogonType)

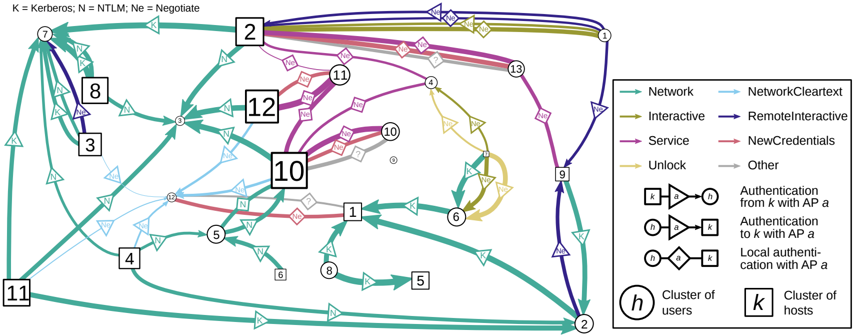
Case 1: $H_{\mathrm{src}} = H_{\mathrm{dst}}$

Case 2: $H_{\mathrm{src}} \neq H_{\mathrm{dst}}$

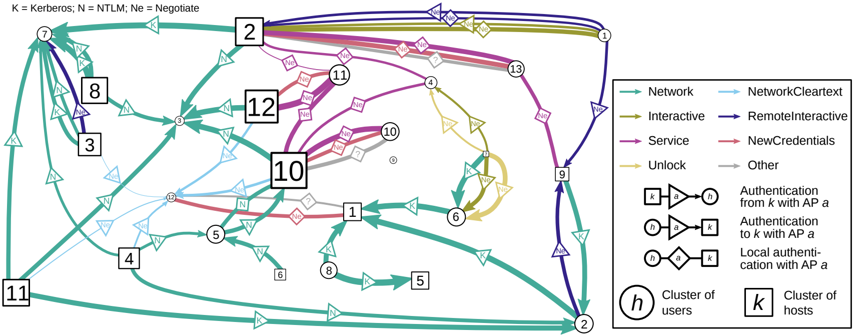


- 74,049 users (top nodes)
- 16,119 hosts (bottom nodes)
- 44 edge types
- 869,547 edges

Second case study – Authentication logs (results)



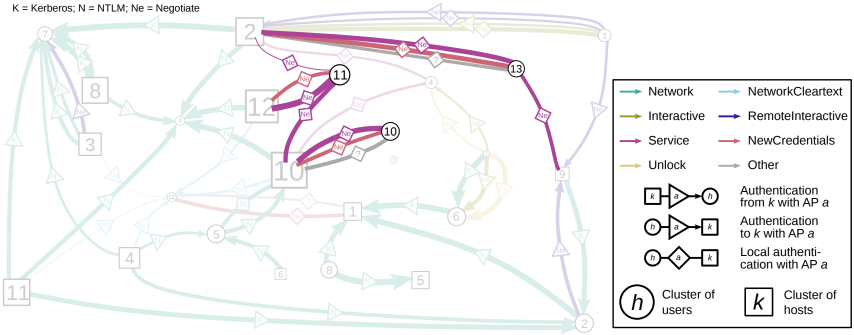
Second case study – Authentication logs (results)



Relevant clusters:

Supicious behaviors:

Second case study – Authentication logs (results)

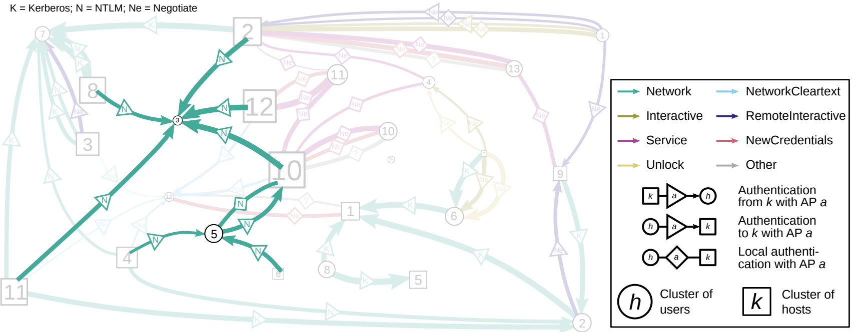


Relevant clusters:

- Service accounts

Supicious behaviors:

Second case study – Authentication logs (results)

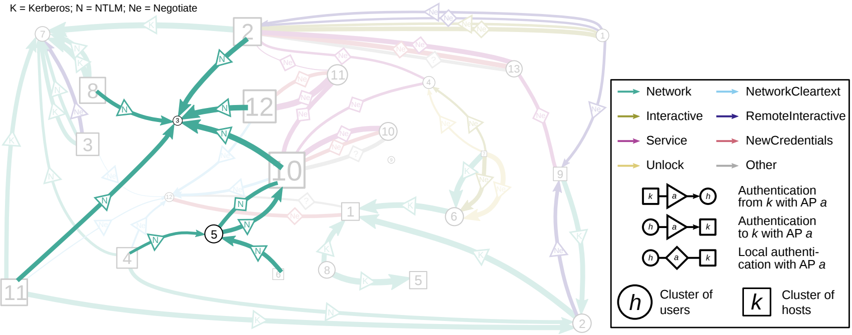


Relevant clusters:

- Service accounts
- Anonymous credentials

Supicious behaviors:

Second case study – Authentication logs (results)



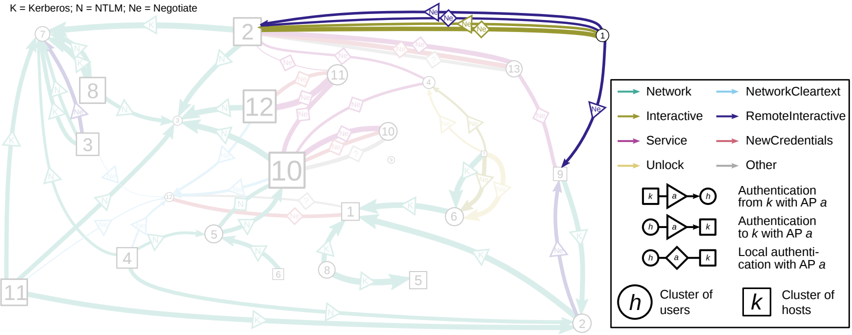
Relevant clusters:

- Service accounts
- Anonymous credentials

Supicious behaviors:

- Compromised user accounts among anonymous credentials

Second case study – Authentication logs (results)



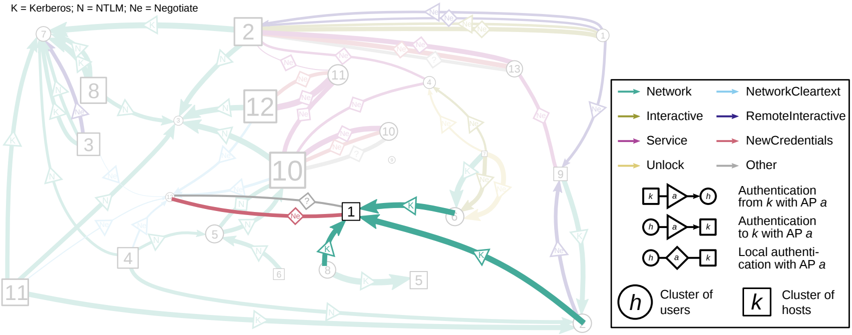
Relevant clusters:

- Service accounts
- Anonymous credentials
- Potential admin accounts

Supicious behaviors:

- Compromised user accounts among anonymous credentials

Second case study – Authentication logs (results)



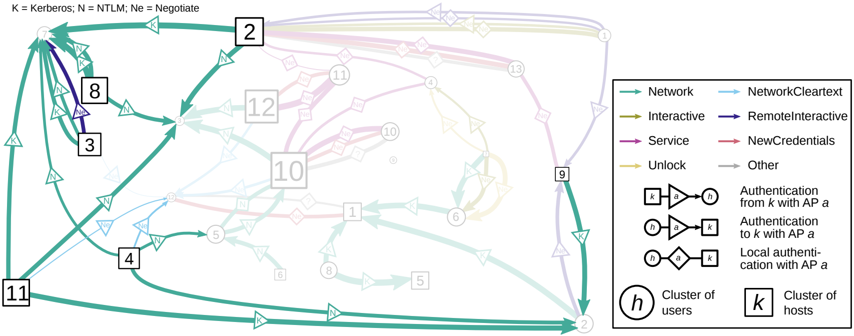
Relevant clusters:

- Service accounts
- Anonymous credentials
- Potential admin accounts
- Servers

Supicious behaviors:

- Compromised user accounts among anonymous credentials

Second case study – Authentication logs (results)



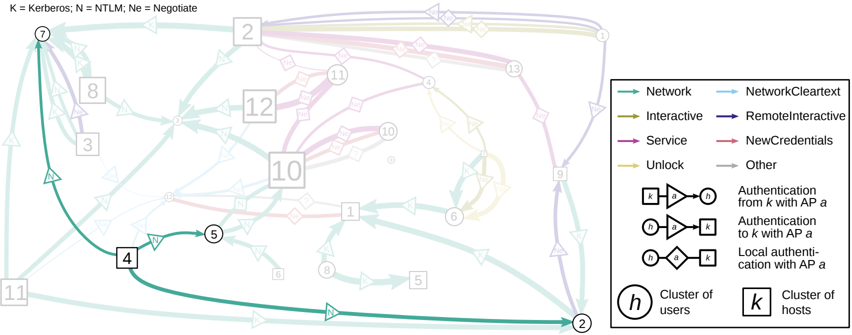
Relevant clusters:

- Service accounts
- Anonymous credentials
- Potential admin accounts
- Servers
- Workstations

Supicious behaviors:

- Compromised user accounts among anonymous credentials

Second case study – Authentication logs (results)



Relevant clusters:

- Service accounts
- Anonymous credentials
- Potential admin accounts
- Servers
- Workstations

Supicious behaviors:

- Compromised user accounts among anonymous credentials
- Outbound NTLM authentications mostly originating from compromised host

Conclusion and perspectives

Contributions

We propose a **graph-oriented approach** to event log exploration. Our method uncovers **meaningful clusters** of entities, and it helps **detect suspicious behaviors**. Overall, it facilitates exploratory analysis by **summarizing** the information contained in the logs.

Future work:

- Better model selection criteria
- Adding a temporal dimension
- Clustering edge types in addition to top and bottom nodes

References

[Ball et al., 2004] Ball, R., Fink, G. A., and North, C. (2004). Home-centric visualization of network traffic for security administration. In *VizSec/DMSec*.

[Biernacki et al., 2000] Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(7):719–725.

[Govaert and Nadif, 2010] Govaert, G. and Nadif, M. (2010). Latent block model for contingency table. *Commun. Stat. Theory Methods*, 39(3):416–425.

[Siadati et al., 2016] Siadati, H., Saket, B., and Memon, N. (2016). Detecting malicious logins in enterprise networks using visualization. In *VizSec*.

[Taylor et al., 2009] Taylor, T., Paterson, D., Glanfield, J., Gates, C., Brooks, S., and McHugh, J. (2009). Flovis: Flow visualization system. In *CATCH*.

[Tomonaga, 2017] Tomonaga, S. (2017). Visualise event logs to identify compromised accounts - logontracer.