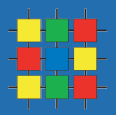
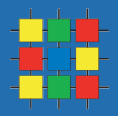


Detekce anomalií a její aplikace v síťové bezpečnosti

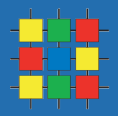
Tomáš Pevný

Department of Computers, Czech Technical University

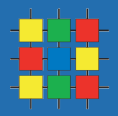
23rd May 2012



Outline



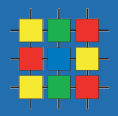
Outline



Motivation

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior.

V. Chandola, A. Banerjee, and V. Kumar, Anomaly detection: a survey, 2009



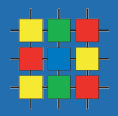
Taxonomy

Unsupervised

- data are presented in a batch with nominal and outliers being mixed.
- does not have a distinct training and testing phase.

Supervised

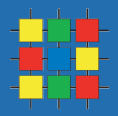
- data from the nominal class are available.
- does have a distinct training and testing phase.



K-nearest neighbor — motivation

Outliers are located far from the nominal data.

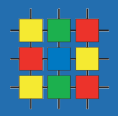
S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, 2000



K-nearest neighbor — calculation

1. For sample $\{x_i\}_{i=1}^{N}$ calculate its distance to k^{th} nearest neighbor.
2. Return fraction p of samples as outliers.

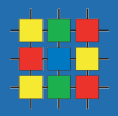
S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, 2000



Local outlier factor — motivation

Outliers have low density with respect to its k neighborhood.

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, Lof: Identifying density-based local outliers, 2000.

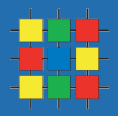


Local outlier factor — calculation

1. For every $\{x_i\}_{i=1}^N$ estimate the local density, $ld_k(x_i)$, as an inverse of average robust distance to k nearest neighbor.
2. Compare density of x_i with that of its k nearest neighbors, P_k,

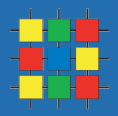
$$lof_k(x_i) = \frac{1}{k} \sum_{x \in P_k} \frac{ld_k(x)}{ld_k(x_i)}.$$

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, Lof: Identifying density-based local outliers, 2000.



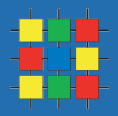
Other approaches

- S. Papadimitriou, H. Kitagawa, P.B. Gibbons, C. Faloutsos, LOCI: fast outlier detection using the local correlation integral, 2003



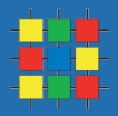
Model-based anomaly detection

Data follows a known distribution, which is (robustly fit) from observed data.



Model-based anomaly detection

- Gaussian distribution $N(\mu, \Sigma)$:
  - Σ is diagonal — independent features,
  - Σ is not diagonal — dependent features.
- Mixture of Gaussian distributions.



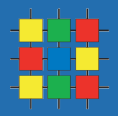
Parzen window estimator

The density in point x is estimated from training points $\{x_i\}_{i=1}^{N}$ as

$$f(x) = \frac{1}{hN} \sum_{i=1}^{N} k\left(\frac{x - x_i}{h}\right),$$

where k is the kernel (e.g. Gaussian kernel).

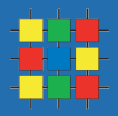
E. Parzen, On Estimation of a Probability Density Function and Mode, 1962



One-class support vector machines

Estimates the support of the probability distribution allowing at most ν false positive rate.

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. Smola, R. C. Williamson, Estimating the support of a high-dimensional distribution, 2001



One-class support vector machines

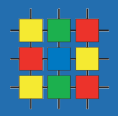
Finds the hyper-plane separating the data from the origin with the highest margin, allowing at most ν misclassified points.

$$\arg \min_{w \in \mathbb{R}^d, \rho} \frac{1}{2}\|w\|^2 - \rho + \frac{1}{\nu N}\sum_{i=1}^{N} \xi_i$$

subject to

$$\begin{aligned} \langle w, x_i \rangle &\geq \rho - \xi_i \\ \xi_i &\geq 0. \end{aligned}$$

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. Smola, R. C. Williamson, Estimating the support of a high-dimensional distribution, 2001



One-class support vector machines

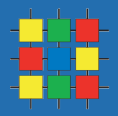
$$\arg \min_{w \in \mathscr{H}, \rho} \frac{1}{2}\|w\|^2 - \rho + \frac{1}{\nu N}\sum_{i=1}^{N} \xi_i$$

subject to

$$\begin{aligned} \langle w, k(x_i, .)\rangle_{\mathscr{H}} &\geq \rho - \xi_i \\ \xi_i &\geq 0. \end{aligned}$$

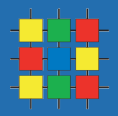
$k(x_i, .)$ is a feature map induced by the chosen kernel.

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. Smola, R. C. Williamson, Estimating the support of a high-dimensional distribution, 2001



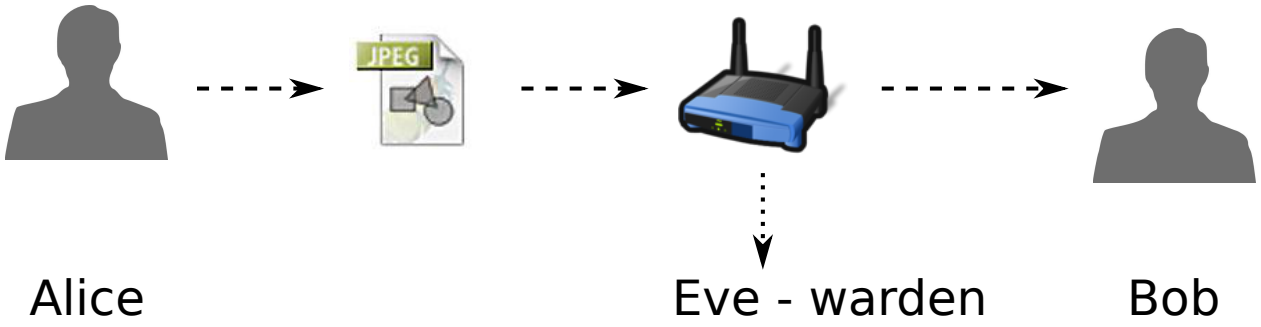
Other approaches

- C. Bennett, K. P. Campbell, A linear programming approach to novelty detection, 2001
- D. M. J. Tax, R. P. W. Duin, Support vector data description, 2004
- K. Noto, C. Brodley, D. Slonim, FRaC: Feature-modeling approach for semi-supervised and unsupervised anomaly detection, 2012



Outline

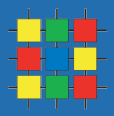
Steganographic channel



Example — LSB replacement

Cover image

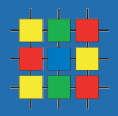
Message



Example — LSB replacement

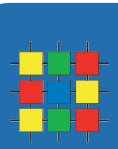
Cover image

Stego image carrying a message



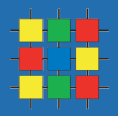
Who uses steganography

- Secret services
- Dissidents
- Communication of botnet with Command and Control channel



Traditional approach to steganalysis

- Set of training cover images and stego images.
- Extract steganalytic features from the training images.
- Train classifier to recognize covers from stegos.

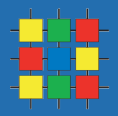


Motivation

Steganalyzers detects only

- known steganographic algorithm,
- known length of the message has to be known,
- known source of covers.

What happen when Eve does not know all details?



Motivation

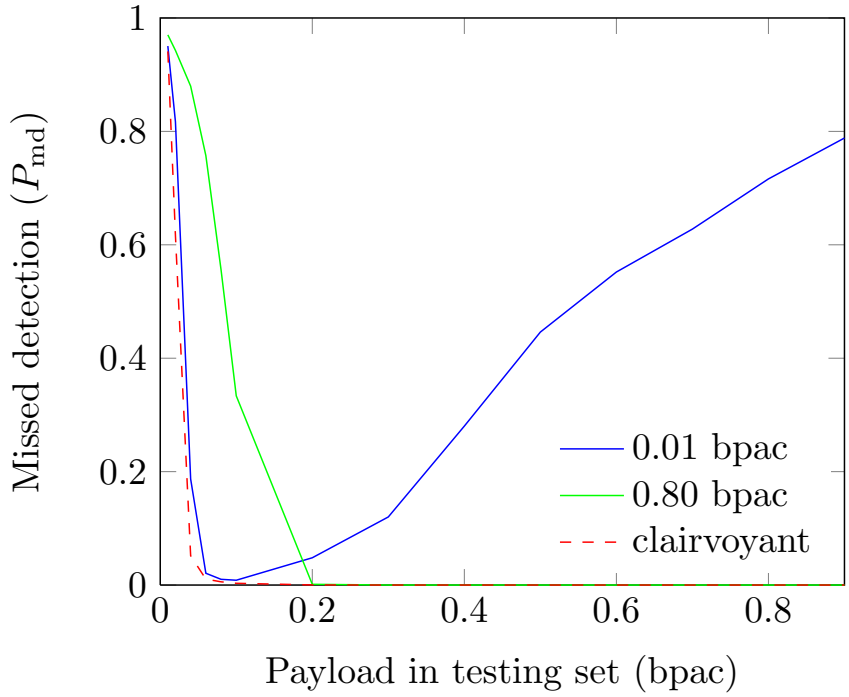
Steganalyzers detects only

- known steganographic algorithm,
- known length of the message has to be known,
- known source of covers.

What happen when Eve does not know all details?

Unknown payload

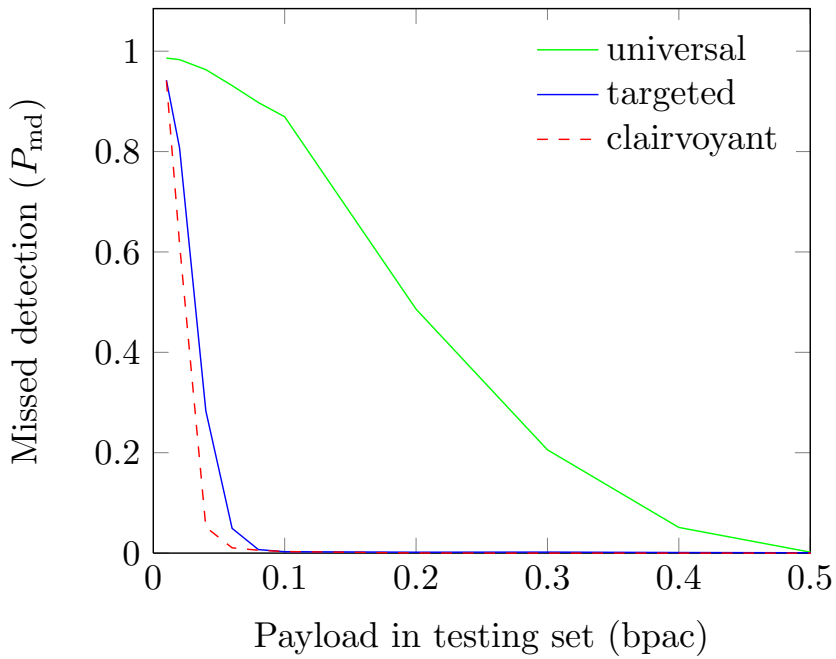
- Detectors trained on images with a fixed embedding rate 0.01 and 0.80 bpac.
- *Clairvoyant* detector knows the possible payload.



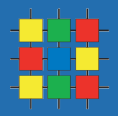
T. Pevný, Detecting messages of unknown length, 2011.

Unknown steganographic algorithms

- *Universal* detector does not know the algorithm.
- *Targeted* detector does not know the payload.
- *Clairvoyant* knows everything.



T. Pevný, Novelty detection in blind steganalysis, 2008.



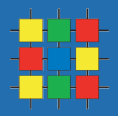
Unknown source of covers

bpp	cover source	CAMERA	BOWS2	JPEG85	NRCS
0.25	match	0.05	0.05	0.01	0.17
	mismatch	0.34	0.17	0.32	0.39

Tab: Probability of error $P_{\text{err}} = 0.5(P_{\text{fp}} + P_{\text{fn}})$

- **match** trained and tested on the same database of images.
- **mismatch** trained on other databases then tested.

T. Pevný, P. Bas, and J. Fridrich, Steganalysis by subtractive pixel adjacency matrix, 2010

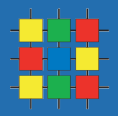


Summary

State of the art steganalyzers achieve great performance.

They can dreadfully fail due to:

- over-fitting to a particular payload,
- over-fitting to certain class of images (cover source),
- missing knowledge of the attacked algorithm.



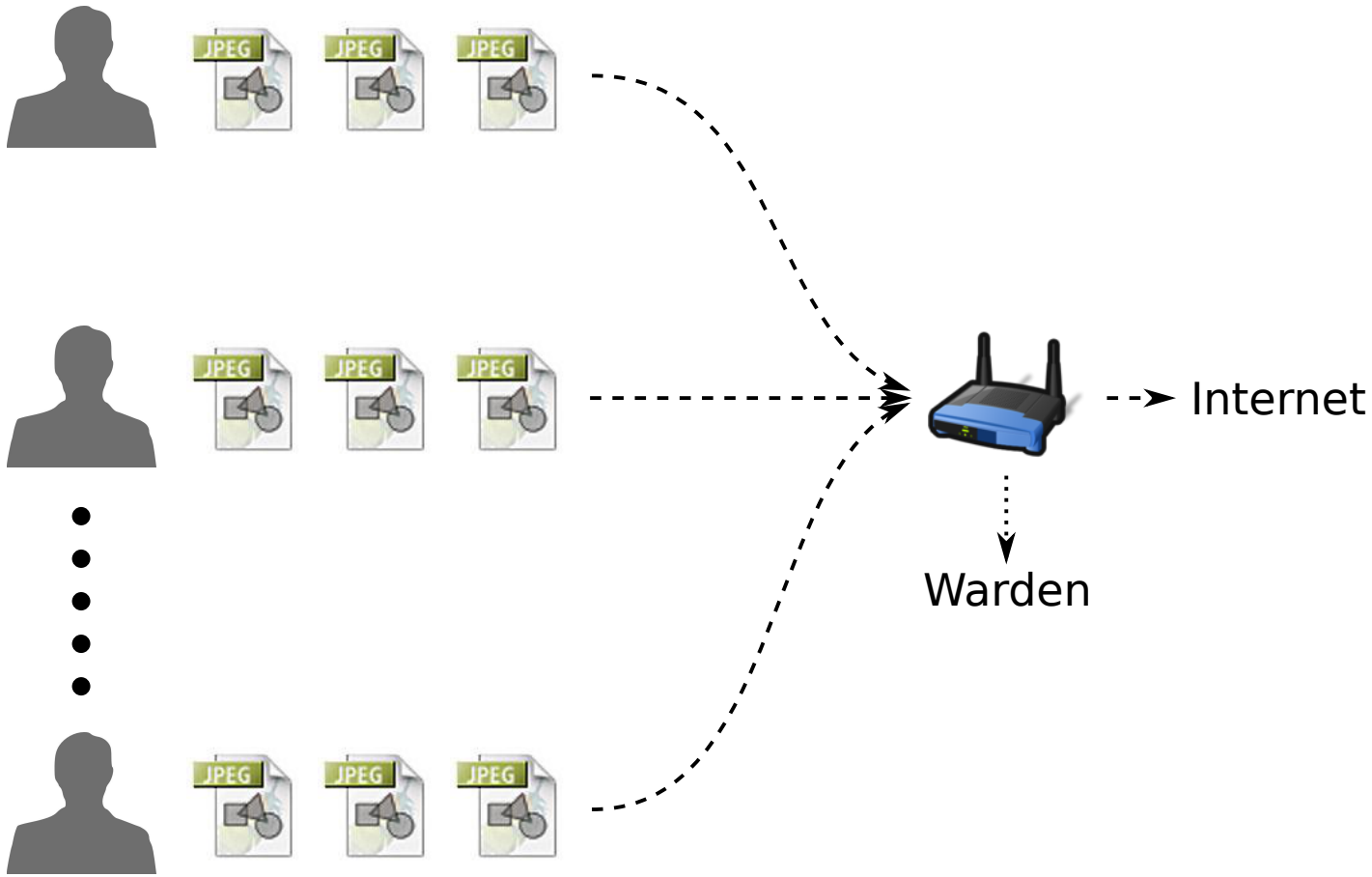
Summary

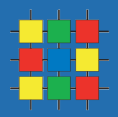
State of the art steganalyzers achieve great performance.

They can dreadfully fail due to:

- over-fitting to a particular payload,
- over-fitting to certain class of images (cover source),
- missing knowledge of the attacked algorithm.

Realistic scenario

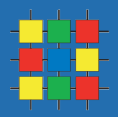




The new paradigm

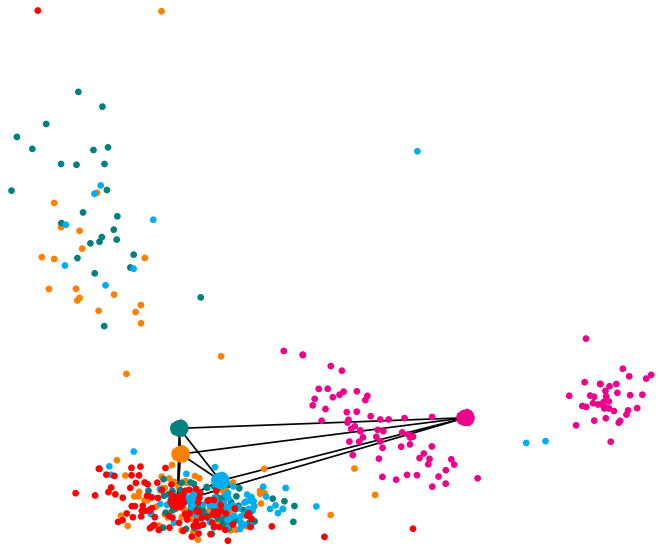
- Extract features.
- Calculate distances between actors (MMD).
- Identify the steganographer(s). local outlier factor (LOF)

A. D. Ker, T. Pevný, Identifying a steganographer in realistic and heterogeneous data sets, 2012



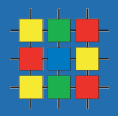
The new paradigm

- Extract features.
- Calculate distances between actors (MMD).
- Identify the steganographer(s). local outlier factor (LOF)



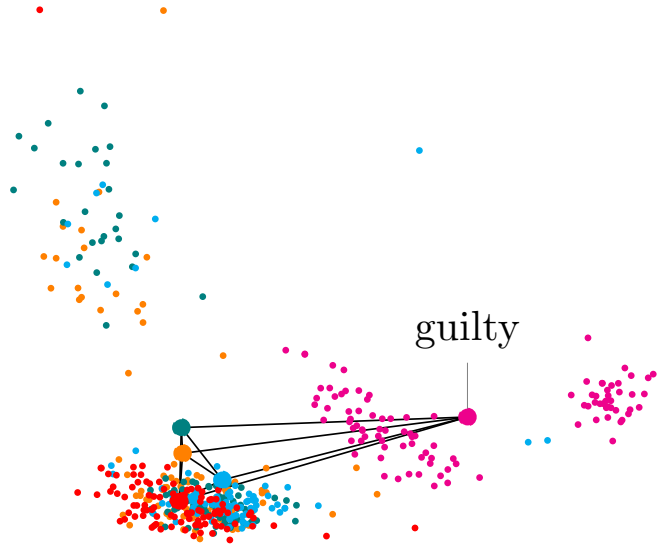
MMD: $D(X, Y) = \sup_f E[f(X)] - E[f(Y)]$.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. Smola, A Kernel Two-Sample Test, 2012

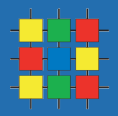


The new paradigm

- Extract features.
- Calculate distances between actors (MMD).
- Identify the steganographer(s). local outlier factor (LOF)



A. D. Ker, T. Pevný, Identifying a steganographer in realistic and heterogeneous data sets, 2012

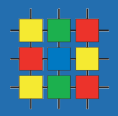


Realistic, heterogenous dataset

On a leading social networking site...

- some users permit global access to images they appear in
- we can click next image or see more of user (if user permits).

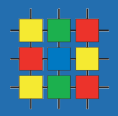
Automated process of following links, restricted to 'Oxford University' users, resulted in 4,051,928 images from 78,107 uploaders.



Realistic, heterogenous dataset

Dataset

- Selected 200 images from each of 4000 uploaders (actors).
- Filtered only for triviality and standard JPEG quality factor.
- Same quality factor, similar size (around 1 Mpix).
- Mixture of sources, even within actors.
- Resampling artifacts.
- Some already-tampered images; hopefully no stego images.



Experiments

- Select 200 images from $n \in \{100, 200, 400, 800, 1600, 3200\}$ actors at random.
- One is the guilty steganographer:
  embeds message using nsF5 in some of their images.
- Perform identification:
  1. Compute PF274 features of every image,
  2. Transform features to whitened space,
  3. MMD distance between each actor,
  4. Rank actors by LOF ($k = 10$).
- Measure how often guilty actor appears in top 10 or top 1% most suspicious.

Accuracy of identification of guilty user

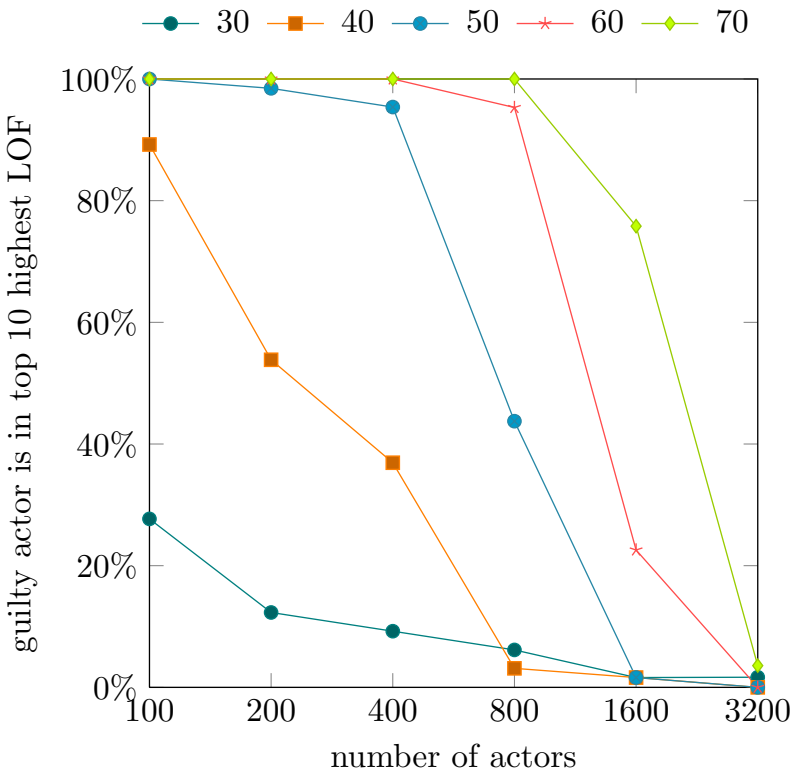


Fig: top 10

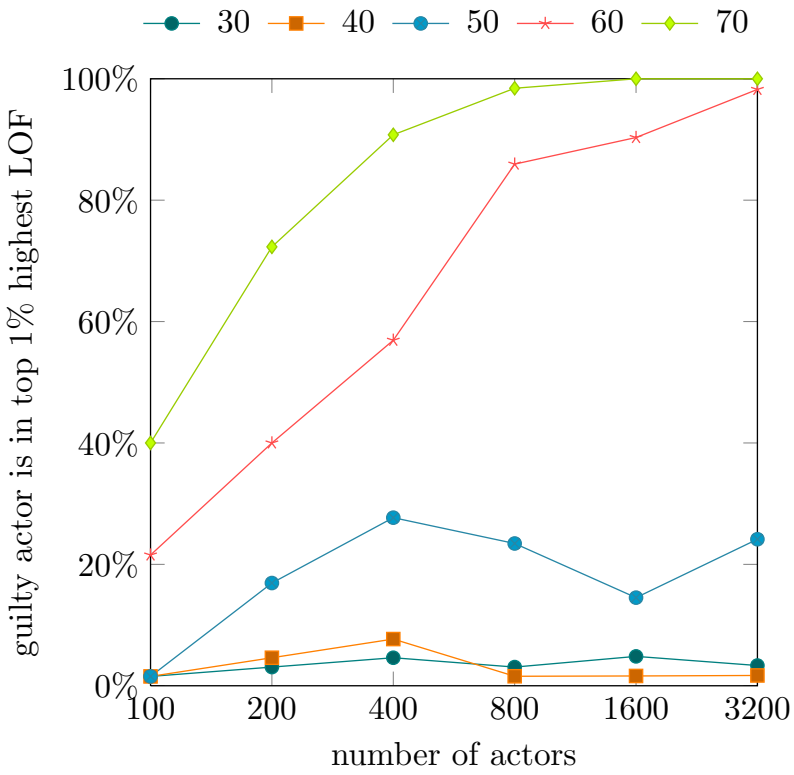
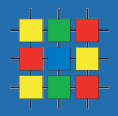


Fig: top 1%



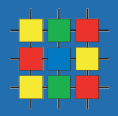
Supervised vs. Unsupervised approach

advantages of the unsupervised approach

- It does not need to be trained.
- It is potentially universal.
- It removes the problem of model mismatch.

limitations of supervised approach

- *Train a classifier on examples of cover and stego images.*
- *The classifier can detect only some steganographic algorithms.*
- *Might have problems with model mismatch.*



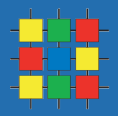
Supervised vs. Unsupervised approach

advantages of the unsupervised approach

- It does not need to be trained.
- It is potentially universal.
- It removes the problem of model mismatch.

limitations of supervised approach

- *Train a classifier on examples of cover and stego images.*
- *The classifier can detect only some steganographic algorithms.*
- *Might have problems with model mismatch.*



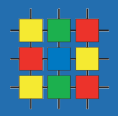
Supervised vs. Unsupervised approach

advantages of the unsupervised approach

- It does not need to be trained.
- It is potentially universal.
- It removes the problem of model mismatch.

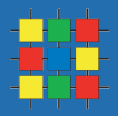
limitations of supervised approach

- *Train a classifier on examples of cover and stego images.*
- *The classifier can detect only some steganographic algorithms.*
- *Might have problems with model mismatch.*

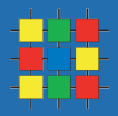


Conclusions

- Identifying steganographer(s) means working on the level of actors, not individual images.
  Allows us to identify a 'typical' level of model mismatch. Make heterogeneous data work for us, not against us.
- This method works on real-world data.
- Potential for universal unsupervised steganalysis...
  ...effectively self-training as long as the majority of actors are innocent.



Outline



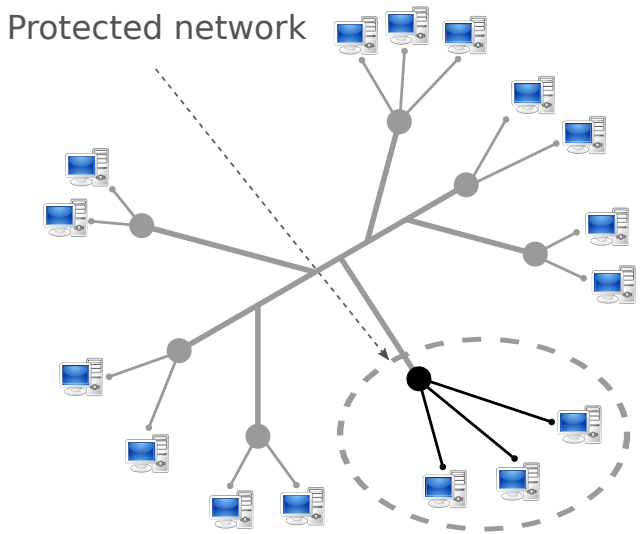
Motivations

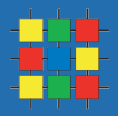
goals:

- to detect attacks,
- to identify suspicious hosts.

approach:

- behavior analysis,
- rather then signature matching.

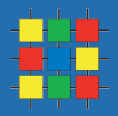




Behavioral analysis

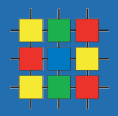
duration	srcIP:port		dstIP:port	flags	packets
3.831	41.58.29.111:1936	→	147.32.80.79:6667	.A.RS.	8
14.859	41.58.29.111:1911	→	147.32.80.79:6667	.A.RS.	14
1.286	41.58.29.111:4205	→	147.32.80.79:6667	.A.RS.	6
1.286	41.58.29.111:4201	→	147.32.80.79:6667	.A.RS.	6
⋮	⋮	→	⋮	⋮	⋮

Sample of netflow data.



The goal

The goal is to model behavior of an ordinary user and identify deviations from this model.

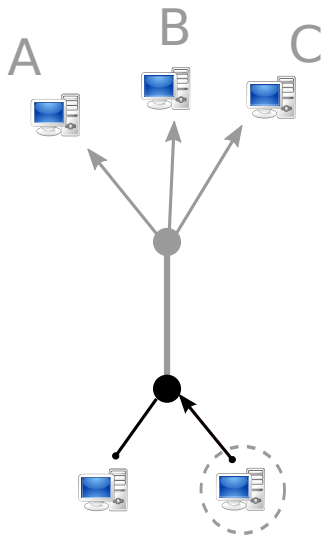


The detector

1. Characterize traffic of every host by a numerical vector.
2. Builds two models by using past data from all observed hosts (incoming and outgoing traffic).
3. Identify anomalies by using two different measures.

A. D. Ker, M. Rehák, M. Grill, Identifying suspicious users in corporate networks, 2012

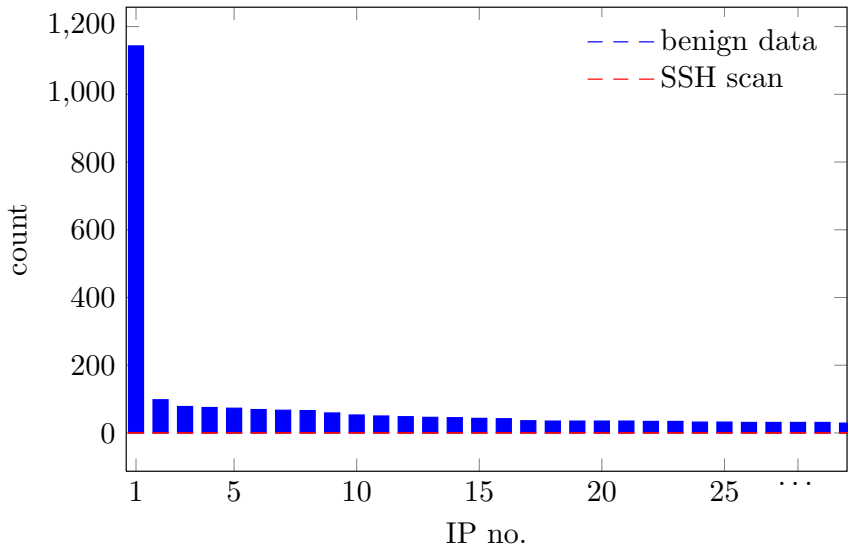
Modeling host's traffic in one time slot



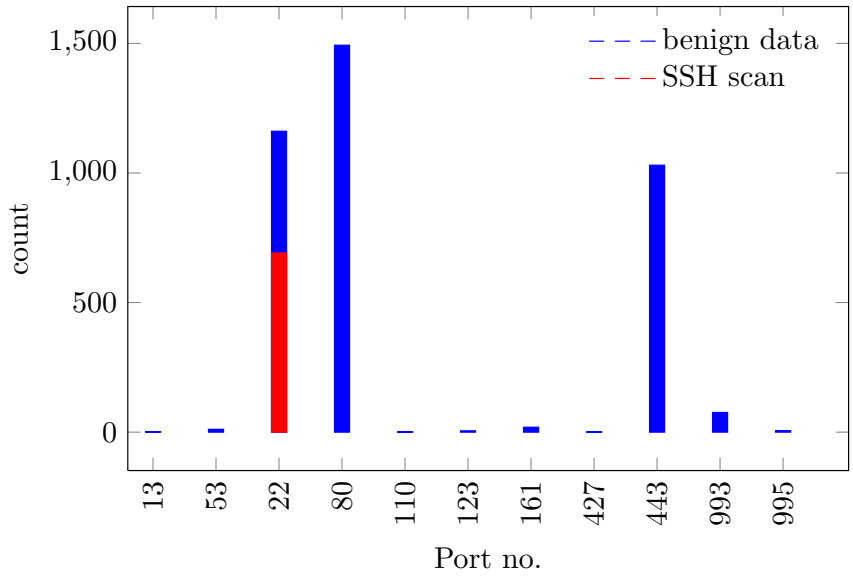
Sample of host's outgoing traffic

modeled host srcIP:port		contacted hosts dstIP:port	
41.58.29.111:1936	→	147.32.80.79:6667	A
41.58.29.111:1911	→	129.39.47.109:53	C
41.58.29.111:4205	→	147.32.80.79:6667	A
41.58.29.111:4201	→	128.226.80.32:8080	B
41.58.29.111:1911	→	147.32.80.79:6667	A
⋮	→	⋮	

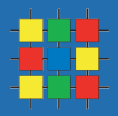
Examples of distributions



Histogram of destination IPs



Histogram of destination ports



Correlation of quantities

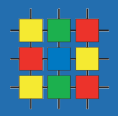
	$t-4$	$t-3$	$t-2$	$t-1$	t
H_{dIP}	0.74	0.76	0.76	0.80	1
H_{dPr}	0.70	0.71	0.72	0.78	1
H_{sPr}	0.41	0.44	0.49	0.56	1

Correlation with respect to time

	H_{dIP}	H_{dPr}	H_{sPr}
H_{dIP}	1	0.70	-0.85
H_{dPr}	0.70	1	-0.95
H_{sPr}	-0.85	-0.95	1

Correlation in one time window

- H^{τ}_{sPr} entropy of distribution of source ports,
- H^{τ}_{dPr} entropy of distribution of destination ports,
- H^{τ}_{dIP} entropy of distribution of destination IP addresses.



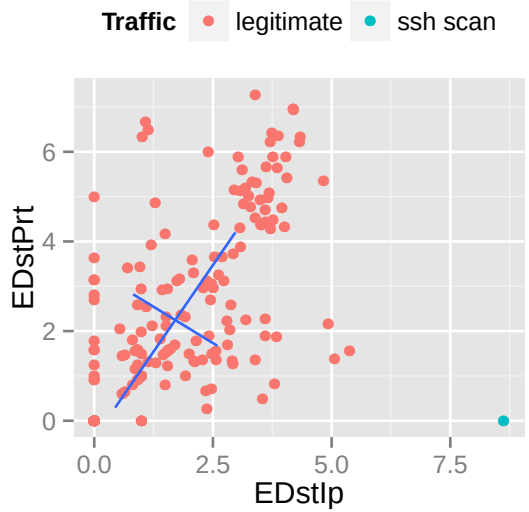
Modeling host's outgoing traffic

Emphasize here the correlation among the variables.

$$x^t = (\underbrace{\mathrm{H}^{\tau-4}_{\mathrm{sPr}}, \mathrm{H}^{\tau-4}_{\mathrm{dPr}}, \mathrm{H}^{\tau-4}_{\mathrm{dIP}}}_{\text{time window } \tau-4}, \underbrace{\mathrm{H}^{\tau-3}_{\mathrm{sPr}}, \mathrm{H}^{\tau-3}_{\mathrm{dPr}}, \mathrm{H}^{\tau-3}_{\mathrm{dIP}}}_{\tau-3}, \ldots, \ldots, \underbrace{\mathrm{H}^{\tau}_{\mathrm{sPr}}, \mathrm{H}^{\tau}_{\mathrm{dPr}}, \mathrm{H}^{\tau}_{\mathrm{dIP}}}_{\tau}) \in \mathbb{R}^{15}.$$

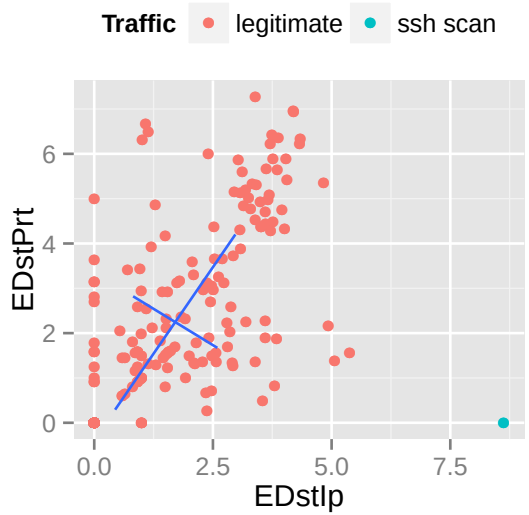
$\mathrm{H}^{\tau}_{\mathrm{sPr}}$	entropy of distribution of source ports,
$\mathrm{H}^{\tau}_{\mathrm{dPr}}$	entropy of distribution of destination ports,
$\mathrm{H}^{\tau}_{\mathrm{dIP}}$	entropy of distribution of destination IP addresses.

Modeling the correlation



$$\left.\begin{array}{rcl} \rightarrow & x_1 \\ \rightarrow & x_2 \\ \vdots & \\ \rightarrow & x_N \end{array}\right\} \mathbf{X} \in \mathbb{R}^{N,15} \qquad \mathbf{X} \xrightarrow{\text{PCT}} \begin{cases} \text{eigenvectors: } y_1, y_2, \ldots, y_r \\ \text{eigenvalues: } \lambda_1 \geq \lambda_r \geq \ldots \geq \lambda_r \end{cases}$$

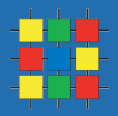
Quantifying the deviation



$$f(x) = \sum_{j=1}^{k} \frac{(y_j^T x)^2}{\lambda_j} \quad \text{variance on major components}$$

$$f^{\perp}(x) = \sum_{j=k+1}^{r} \frac{(y_j^T x)^2}{\lambda_j} \quad \text{variance on minor components}^1$$

[1] Shyu et al., A novel anomaly detection scheme based on principal component classifier, 2003



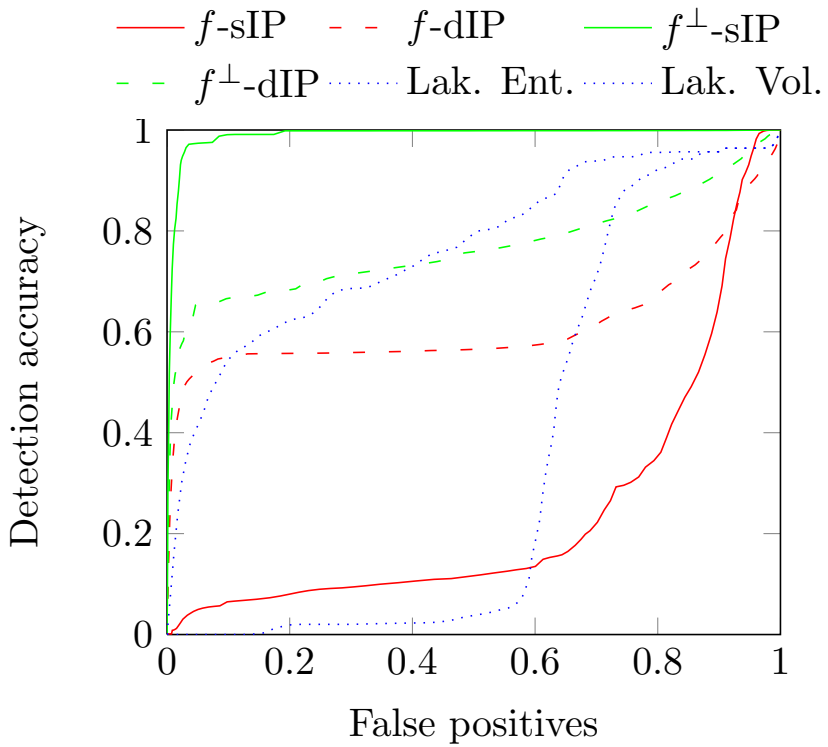
Experimental comparison

- one week long traffic from the university network
- manually labeled by an experienced operator
- compare to the adapted detectors[2] of Lakhina et al.[3]

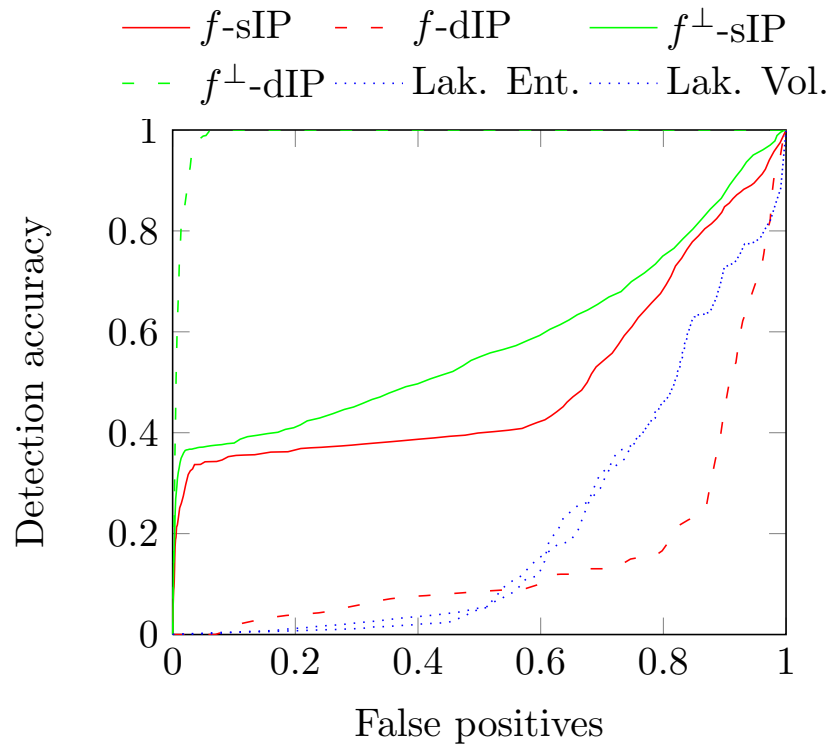
[2] Rehák et al., Network Intrusion Detection by Means of Community of Trusting Agents, 2007

[3] Lakhina et al, Mining Anomalies using Traffic Feature Distributions, 2005

Horizontal scan

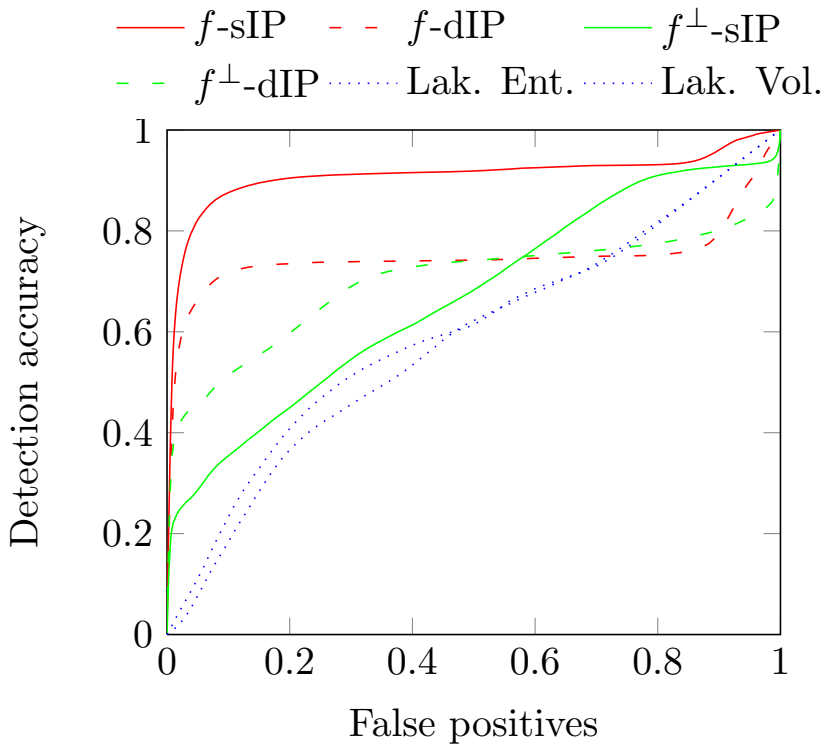


Requests

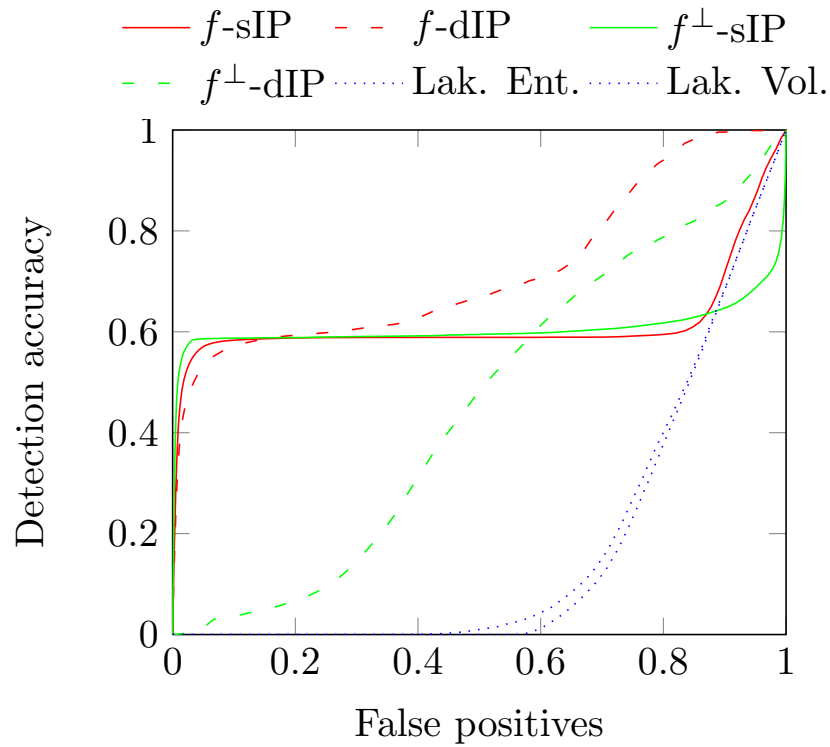


Responses

P2P traffic

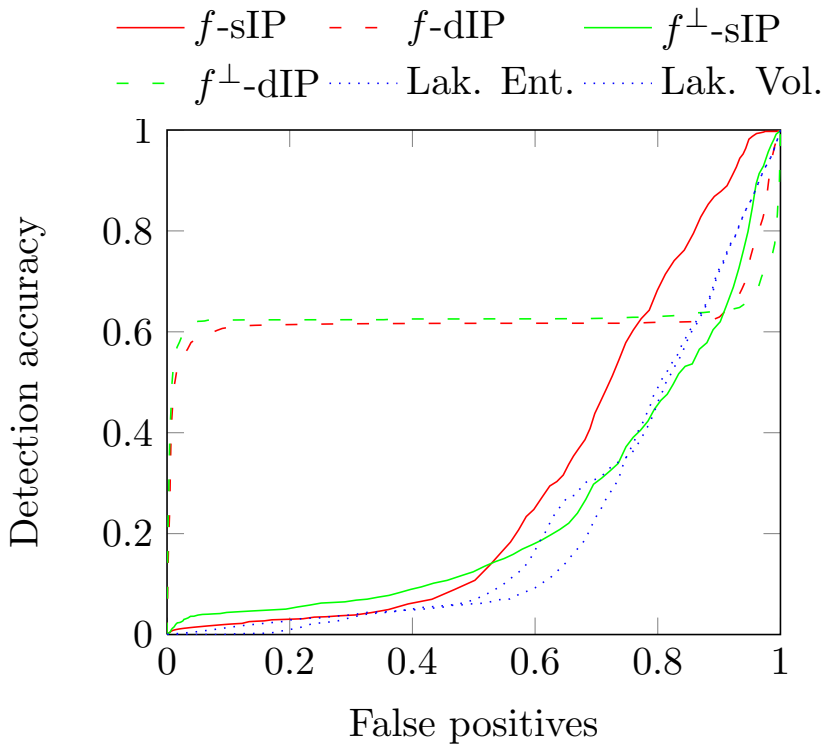


Requests

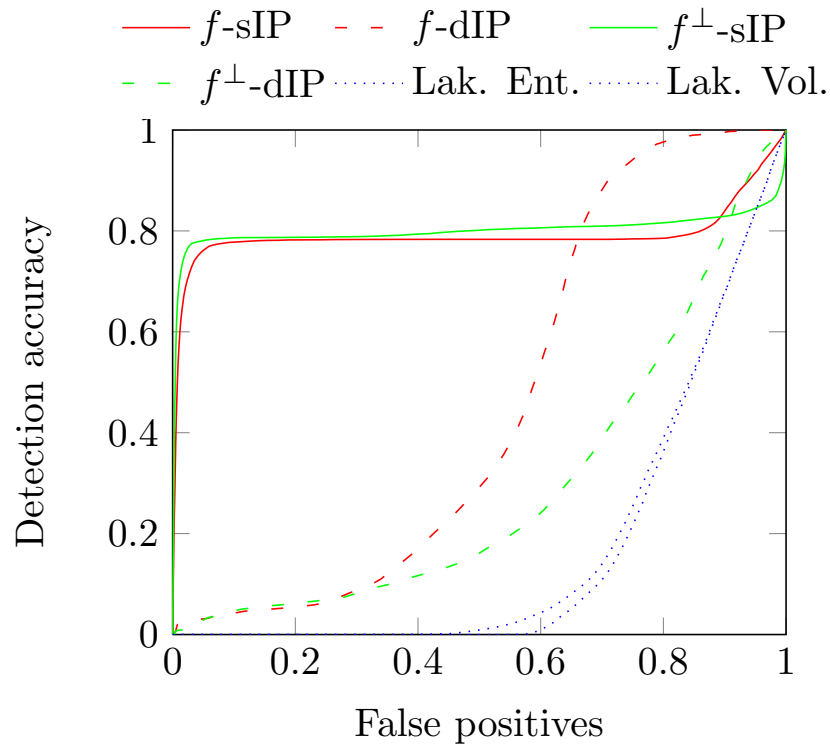


Responses

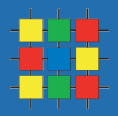
Skype traffic



Requests



Responses



Conclusions

- We presented set of four detectors modeling behavior of users.
- Exploit simultaneously time and spatial correlation.
- Low computational complexity.
- Suspected users usually involved in malicious(unwanted) behavior.