Statistical model for simulation of normal user traffic

FloCon 2015

Jan Stiborek

January 2015

Traditional Network Security

- Traditional network security techniques becomes insufficient
  - Protected perimeter is not strictly defined
  - Not all devices are under direct control (BYOD)
  - Attacks come from inside of network
- Novel attack targeted against network infrastructure

Intrusion Detection System

- Traditionally deep packet inspection – Snort, Bro
- Drawbacks:
  - Novel attacks – need of periodical updates
  - Encrypted traffic
  - High speed networks

Anomaly detection IDS system

- Searches for anomalies in the traffic
- Independent of known attacks database
- No patterns required
- Ability to detect new attacks – “zero day attacks”
- Does not work with actual content (minimal privacy issues, high speed networks)
  - Uses NetFlow/IPFIX data

NetFlow/IPFIX Data Example

Date flow start	Duration	Proto	Src IP (Addr:Port)		Dst IP (Addr:Port)	Flags	Packets	Bytes
12/11/2013 11:58:52.161	0.000	UDP	147.32.80.9:53	->	147.32.86.17:56090	......	4	1832
12/11/2013 11:58:53.459	0.000	UDP	147.32.80.9:53	->	147.32.81.223:53157	......	2	254
12/11/2013 11:58:52.469	0.000	UDP	68.142.254.15:53	->	147.32.80.9:51591	......	2	266
12/11/2013 11:58:54.519	0.000	ICMP	147.32.87.98:3	->	109.169.221.65:1	......	2	152
12/11/2013 11:58:52.408	0.000	UDP	147.32.80.9:50144	->	213.199.180.53:53	......	2	130
12/11/2013 11:58:52.890	0.000	UDP	147.32.80.9:64966	->	193.108.88.129:53	......	2	162
12/11/2013 11:58:48.435	5.117	TCP	147.32.80.13:3128	->	147.32.86.122:2183	.AP.SF	44	18844
12/11/2013 11:58:56.371	0.000	TCP	147.32.83.216:56113	->	178.63.42.124:428	....S.	2	120

Anomaly detection IDS system

- Precise tuning of internal IDS parameters is required
- Difficulties with the evaluation and comparison of different anomaly detection methods
- Evaluation datasets are difficult to obtain
  - Malicious activity is forbidden by company security policy (no matter how beneficial it can be)
  - Lab networks does not correctly mimic statistics of real network
  - Manual labeling does not scale

Simulation – possible answer

- Simulation of malicious activity vs. simulation of the normal user
- Both required to correctly set parameters of IDS
- We propose three different simulation models with different level of details
  - Random sampling
  - Marginal model
  - Time variant join probability model

Random sampling

- Data generated completely randomly
  - No dependency between features
  - Assumes uniform distribution of individual features
  - Restriction: 0 < #bytes ≤ #packets • 65535
- Easy to implement
- Does not require any training data, no manual tuning
- Used as baseline

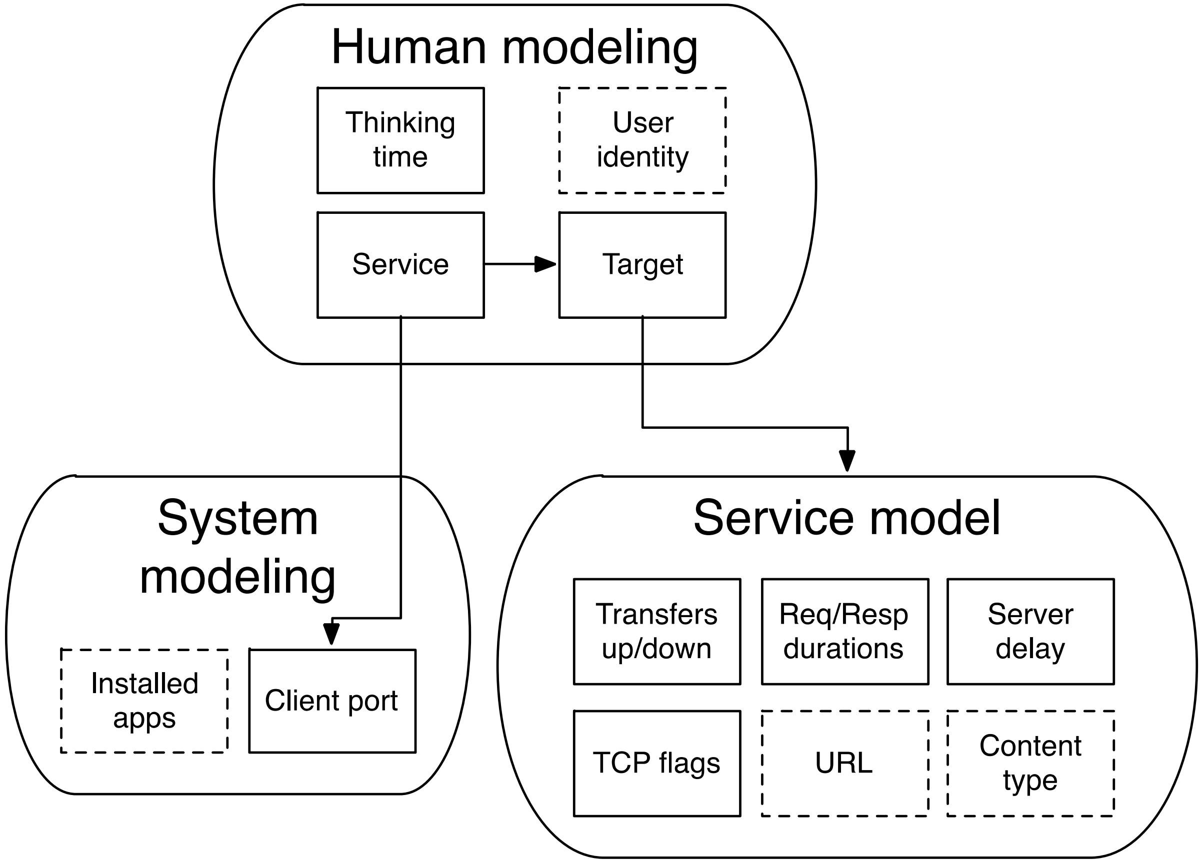
Sampling with independent intra-flow relations — marginal model

- Uses training data to train model of individual NetFlow features
- NetFlows are processed in request/response pairs
- Partially captures inter-flow relations
- NetFlow features modeled independently
  - Non-parametric PDF estimates (Histogram)

Time variant join probability model

- NetFlows are processed in request/response pairs
- Captures more complicated aspects of the user’s behavior missed by previous approaches
  - relations between individual NetFlow features
  - changes of the user’s behavior

Time variant join probability model – structure



- All features depends on the daytime (t)
- The thinking time (T) depends only on the daytime (t)

Time variant join probability model – inner models

- Human modeling

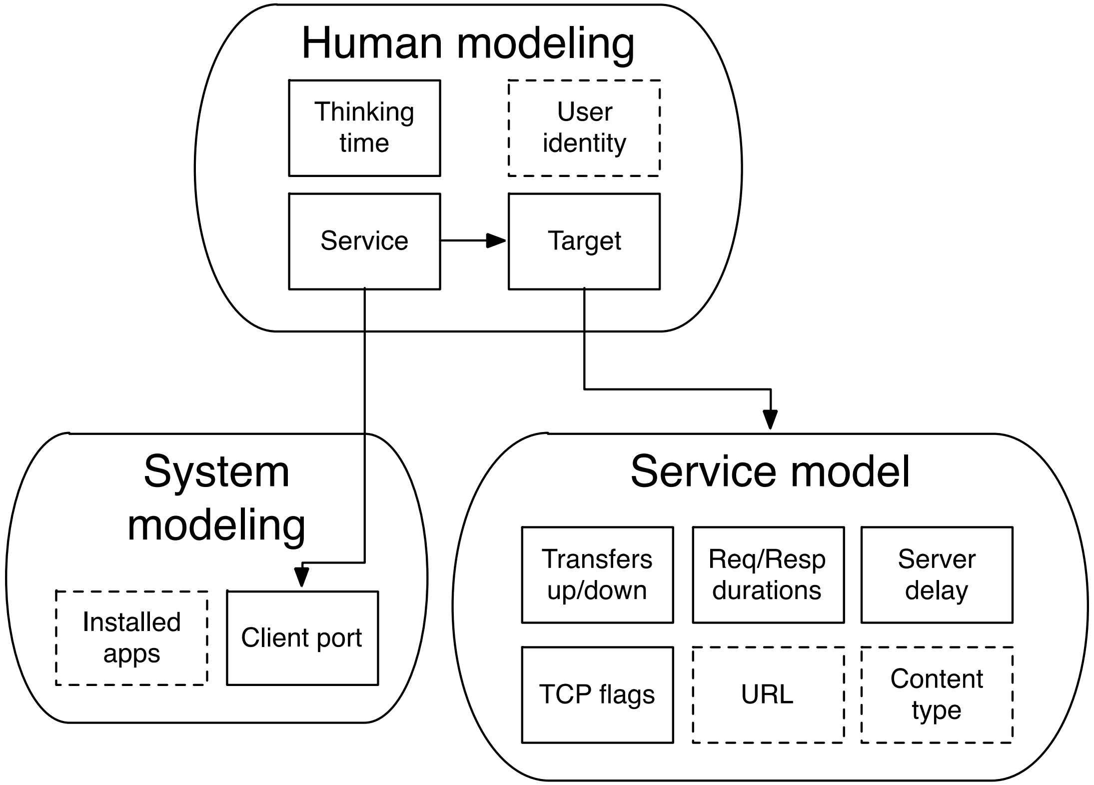
$$p(T \mid t), p(s \mid t), p(dIP \mid s,t)$$

- System modeling

$$p(cPort \mid s,t)$$

- Service modeling

$$p(x_s \mid dIP, s, t)$$



Evaluation – big picture

- Goal is to develop simulation technique that generates realistic traffic for evaluation of AD algorithms
- We measure difference between simulated and real traffic
- We compare results for different simulation techniques and select the optimal one
- If the difference is small, the traffic is realistic enough and it can be used for evaluation

Evaluation – criteria

- Calculated distance between distribution of anomaly scores of real and simulated data
- Used Jensen-Shannon divergence – symmetric and smooth version of Kullback–Leibler divergence

$$JSD(P,Q) = \frac{1}{2}KL(P,M) + \frac{1}{2}KL(Q,M)$$

$$M = \frac{1}{2}(P+Q)$$

Evaluation – detection methods

- Every detection method provides anomaly score in range from 0 (not anomalous at all) to 1 (most anomalous) for every NetFlow
- Selected algorithm:
  - PCA based algorithms: *Pevný-f-dIP, Pevný-f-sIP, Pevný-*$f^{\perp}$*-dIP, Pevný-*$f^{\perp}$*-sIP, Lak.Ent, Lak.Vol.-sIP, Lak.Vol.-dIP*
  - Algorithm with internal model: *Minnesota Intrusion Detection System*
  - Without internal model: *Xu-sIP, Xu-dIP*

Evaluation – selected data

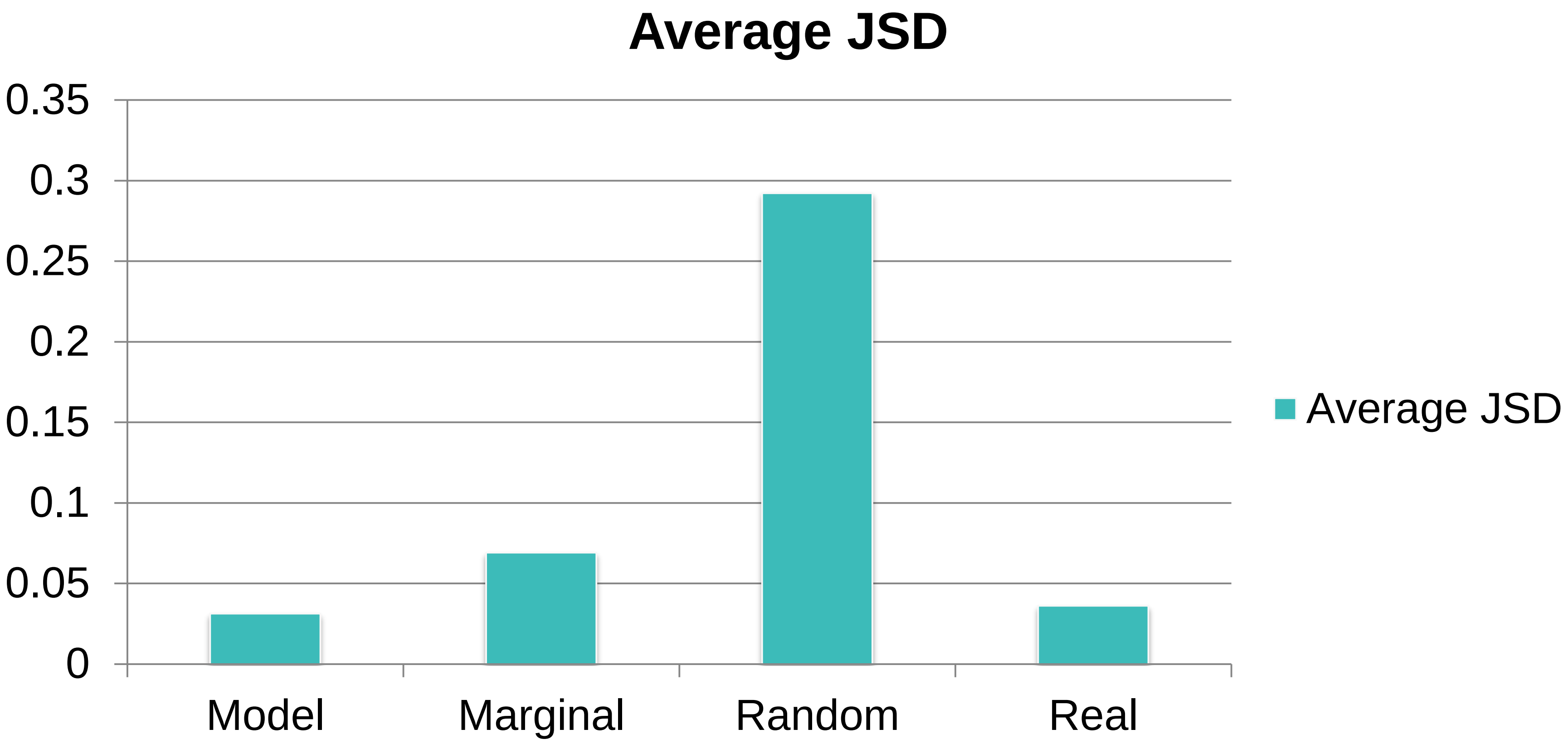
- Data recorded on university campus during the one week in April 2013
- Selected set of full-time employees with various user profiles (developers, scientists, managers and administrative staff)
- Their data were used as training samples for Marginal and Time variant join probability model
- Rest of the traffic served as background traffic

Evaluation – results

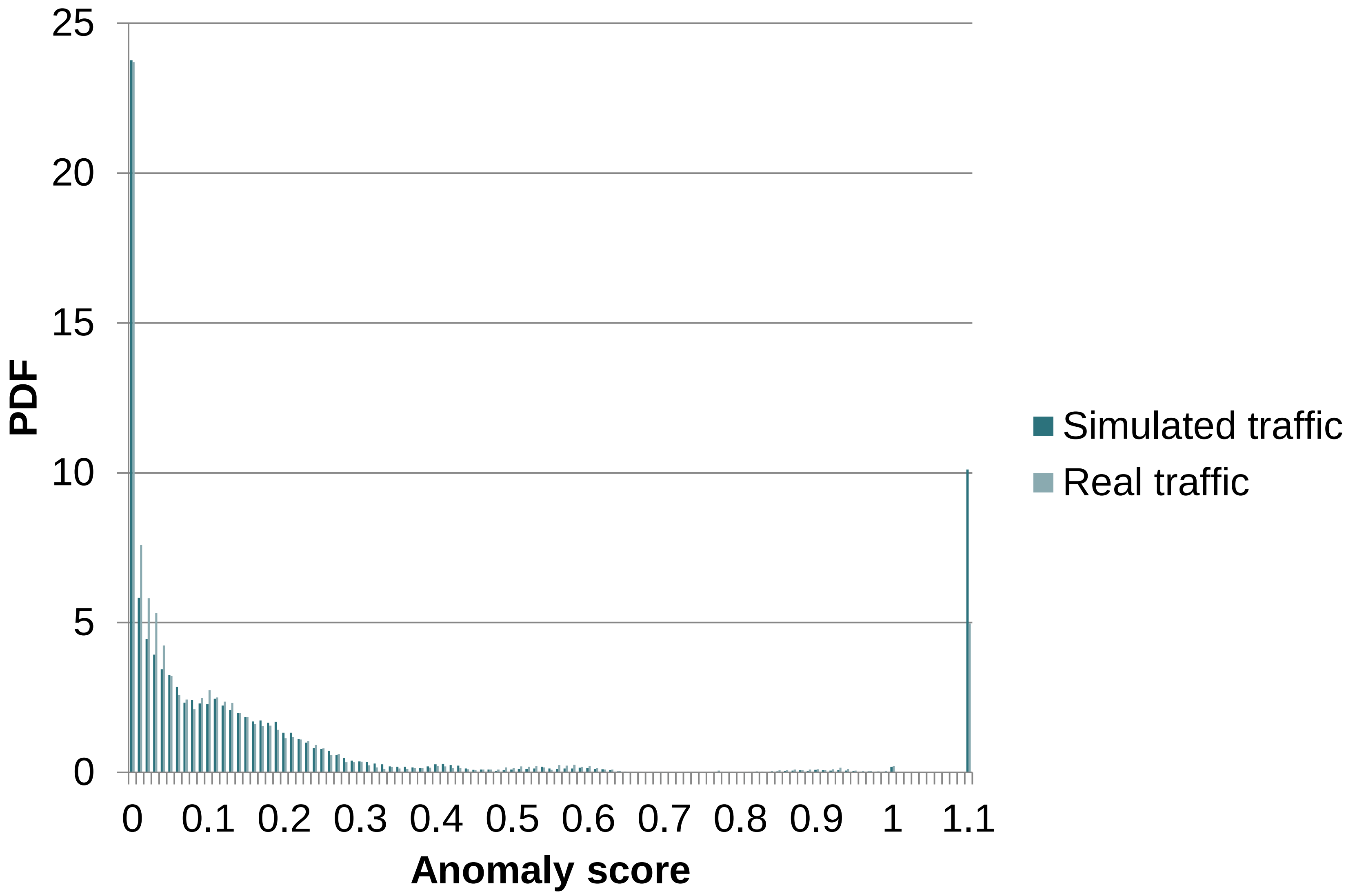
Detection alg.	Model	Marginal	Random	Real
Pevný-f-dIP	**0.0321**	0.0483	0.5427	0.0769
Pevný-f-sIP	**0.0320**	0.0464	0.5573	0.0674
Pevný-$f^{\perp}$-dIP	**0.0124**	0.0214	0.4237	0.0204
Pevný-$f^{\perp}$-sIP	**0.0088**	0.0216	0.3942	0.0198
Lak.Ent.	**0.0472**	0.1111	0.1889	0.0549
Lak.Vol.-sIP	0.0353	0.1132	0.1889	**0.0118**
Lak.Vol.-dIP	0.0433	0.1124	0.1874	**0.0152**
MINDS	**0.0292**	0.0976	0.2399	0.0516
Xu-sIP	0.0301	0.0371	0.0286	**0.0078**
Xu-dIP	0.0421	0.0815	0.1704	**0.0354**
Average	**0.0313**	**0.0691**	**0.2922**	**0.0361**

Jensen-Shannon divergence for distributions of anomaly score for selected AD alg.

Evaluation – results



Evaluation – results



Thank you.