Flow-based network traffic generation using Generative Adversarial Networks*

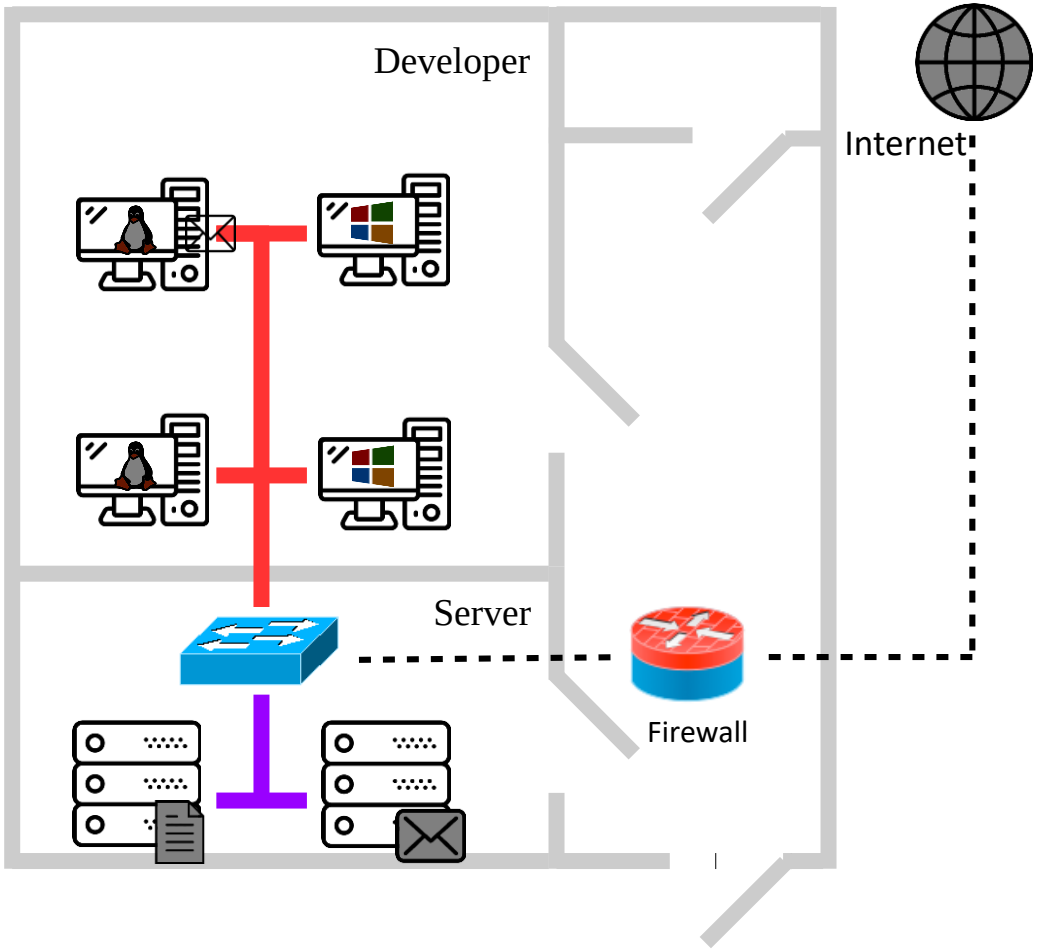
Markus Ring, Daniel Schlör, Dieter Landes and Andreas Hotho

* M. Ring, D. Schlör, D. Landes, A. Hotho: Flow-Based Network Traffic Generation Using Generative Adversarial Networks. In *Computers and Security 82,* 2019, 156-172.

1.) Motivation

HOCHSCHULE COBURG
university of applied sciences

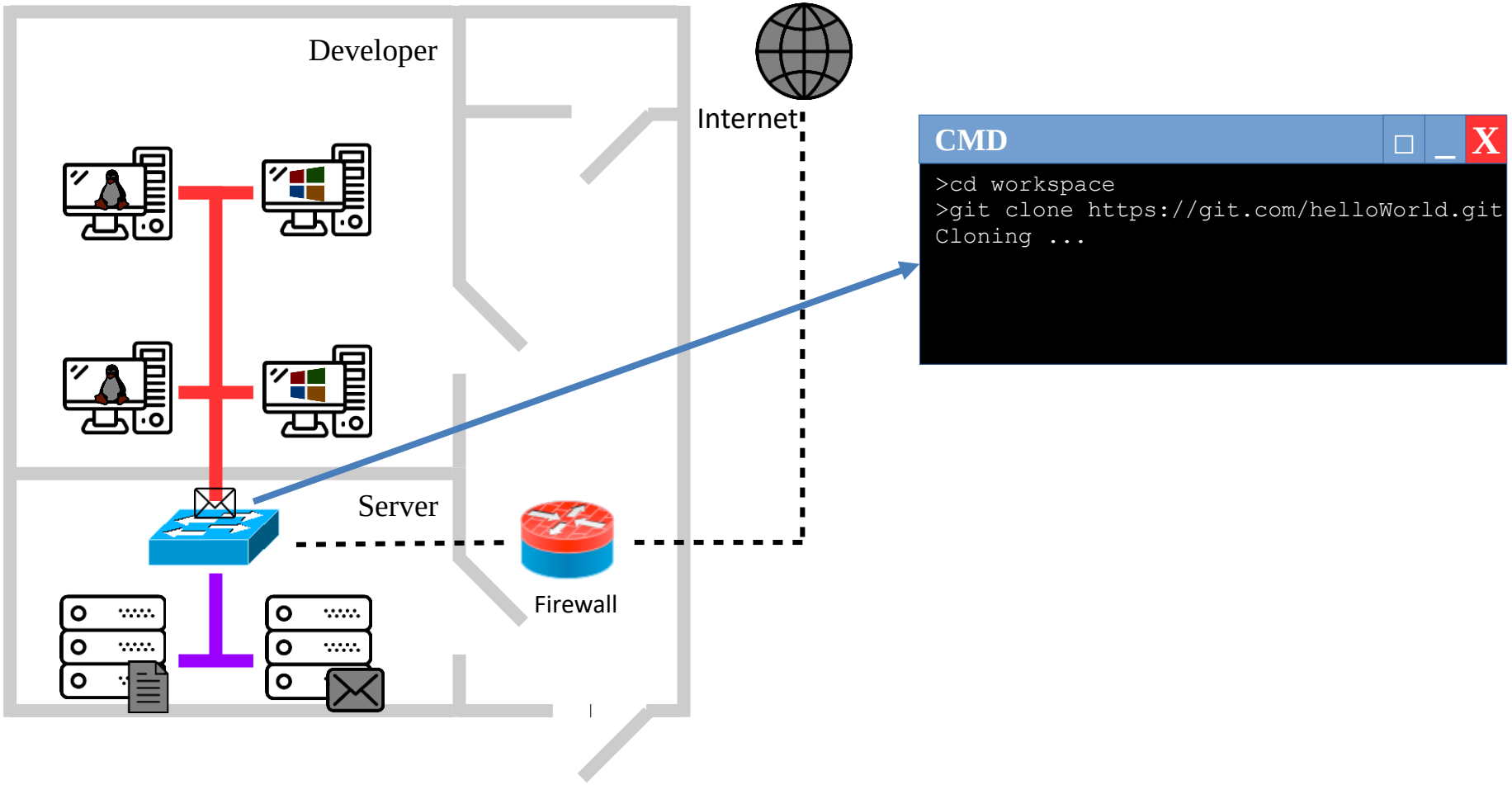
Company network



1.) Motivation

HOCHSCHULE COBURG
university of applied sciences

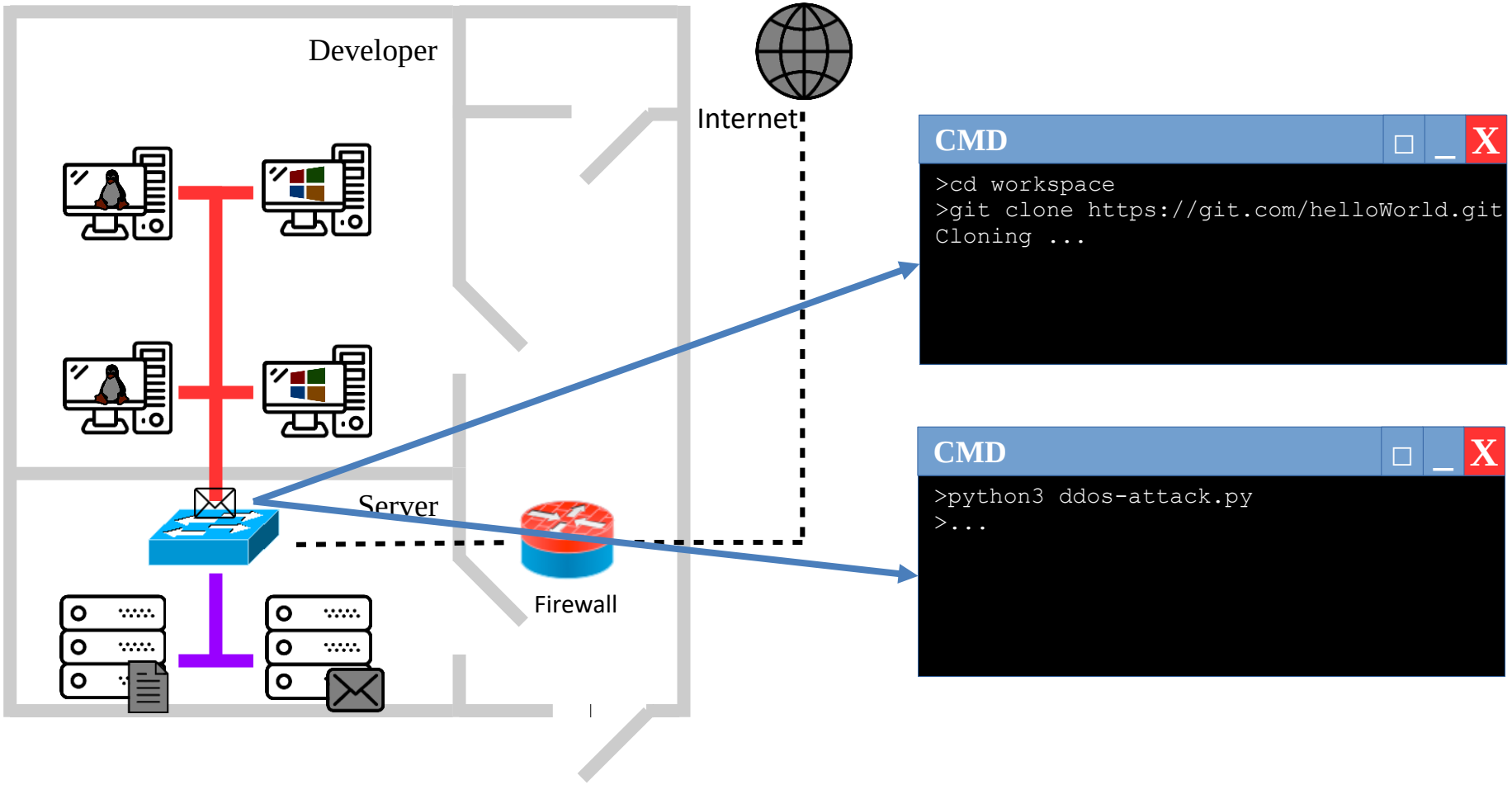
Company network



1.) Motivation

HOCHSCHULE COBURG
university of applied sciences

Company network



1) Motivation

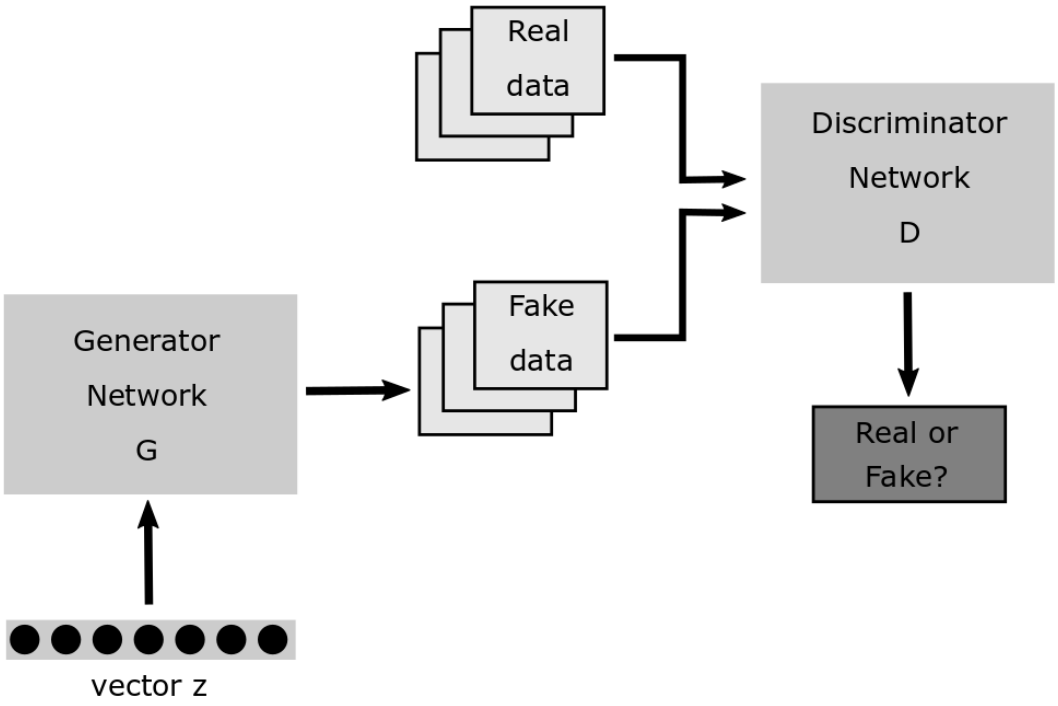
HOCHSCHULE COBURG
university of applied sciences

- **Target:** Generate synthetic network traffic to enrich existing data sources
- **Focus:** Network traffic in flow-based format
  - Flows describe network connections between endpoint devices using meta data

```
2017-03-15 00:01:16.632,    0.000,TCP  ,192.168.100.5,   445,  192.168.220.16, 58844,       1,     108,    1,.AP...,  0
2017-03-15 00:01:16.552,    0.000,TCP  ,192.168.100.5,   445,  192.168.220.15, 48888,       1,     108,    1,.AP...,  0
2017-03-15 00:01:16.551,    0.004,TCP  ,192.168.220.15, 48888,   192.168.100.5,   445,       2,     174,    1,.AP...,  0
2017-03-15 00:01:16.631,    0.004,TCP  ,192.168.220.16, 58844,   192.168.100.5,   445,       2,     174,    1,.AP...,  0
2017-03-15 00:01:16.552,    0.000,TCP  ,192.168.100.5,   445,  192.168.220.15, 48888,       1,     108,    1,.AP...,  0
2017-03-15 00:01:16.631,    0.004,TCP  ,192.168.220.16, 58844,   192.168.100.5,   445,       2,     174,    1,.AP...,  0
2017-03-15 00:01:17.432,    0.000,TCP  ,192.168.220.9, 37884,   192.168.100.5,   445,       1,      66,    1,.A....,  0
2017-03-15 00:01:17.431,    0.000,TCP  ,192.168.100.5,   445,   192.168.220.9, 37884,       1,      70,    1,.AP...,  0
2017-03-15 00:01:17.432,    0.000,TCP  ,192.168.220.9, 37884,   192.168.100.5,   445,       1,      66,    1,.A....,  0
2017-03-15 00:01:17.776,    0.000,TCP  ,23.57.17.35,   443,  192.168.220.16, 45061,       1,     358,    1,.AP...,  0
2017-03-15 00:01:17.782,    0.000,TCP  ,23.57.17.35,   443,  192.168.220.16, 45558,       1,     372,    1,.AP...,  0
2017-03-15 00:01:17.777,    0.000,TCP  ,23.57.17.35,   443,  192.168.220.16, 45487,       1,     357,    1,.AP...,  0
2017-03-15 00:01:17.749,    0.082,TCP  ,23.57.17.35,   443,  192.168.220.16, 45585,       4,     419,    1,.AP.S.,  0
2017-03-15 00:01:17.748,    0.083,TCP  ,23.57.17.35,   443,  192.168.220.16, 45583,       4,     419,    1,.AP.S.,  0
2017-03-15 00:01:17.750,    0.089,TCP  ,23.216.202.232,   443,  192.168.220.16, 51138,       5,     498,    1,.AP.S.,  0
2017-03-15 00:01:17.748,    0.083,TCP  ,23.57.17.35,   443,  192.168.220.16, 45584,       4,     419,    1,.AP.S.,  0
2017-03-15 00:01:17.750,    0.089,TCP  ,23.57.17.35,   443,  192.168.220.16, 45588,       4,     419,    1,.AP.S.,  0
2017-03-15 00:01:17.749,    0.086,TCP  ,23.57.17.35,   443,  192.168.220.16, 45586,       4,     419,    1,.AP.S.,  0
2017-03-15 00:01:17.827,    0.000,UDP  ,192.168.210.4,   138, 192.168.210.255,   138,       1,     243,    1,......,  0
2017-03-15 00:01:17.728,    0.000,UDP  ,192.168.220.16, 35549,    192.129.28.9,    53,       1,      73,    1,......,  0
2017-03-15 00:01:17.728,    0.051,TCP  ,192.168.220.16, 45588,     23.57.17.35,   443,       5,     906,    1,.AP.S.,  0
2017-03-15 00:01:17.760,    0.023,TCP  ,192.168.220.16, 45558,     23.57.17.35,   443,       2,     598,    1,.AP...,  0
2017-03-15 00:01:17.727,    0.044,TCP  ,192.168.220.16, 45584,     23.57.17.35,   443,       5,     906,    1,.AP.S.,  0
2017-03-15 00:01:17.728,    0.047,TCP  ,192.168.220.16, 45586,     23.57.17.35,   443,       5,     906,    1,.AP.S.,  0
2017-03-15 00:01:17.727,    0.048,TCP  ,192.168.220.16, 45585,     23.57.17.35,   443,       5,     906,    1,.AP.S.,  0
2017-03-15 00:01:17.754,    0.021,TCP  ,192.168.220.16, 45061,     23.57.17.35,   443,       2,     641,    1,.AP...,  0
2017-03-15 00:01:17.727,    0.044,TCP  ,192.168.220.16, 45583,     23.57.17.35,   443,       5,     906,    1,.AP.S.,  0
2017-03-15 00:01:17.755,    0.024,TCP  ,192.168.220.16, 45487,     23.57.17.35,   443,       2,     633,    1,.AP...,  0
2017-03-15 00:01:17.728,    0.051,TCP  ,192.168.220.16, 51138,  23.216.202.232,   443,       6,     972,    1,.AP.S.,  0
2017-03-15 00:01:18.403,    0.000,UDP  ,192.129.28.9,    53,  192.168.220.16, 32777,       2,     533,    1,......,  0
2017-03-15 00:01:18.511,    0.003,TCP  ,17.173.65.113,   443,  192.168.220.16, 49062,       2,    1089,    1,.AP..., 32
```

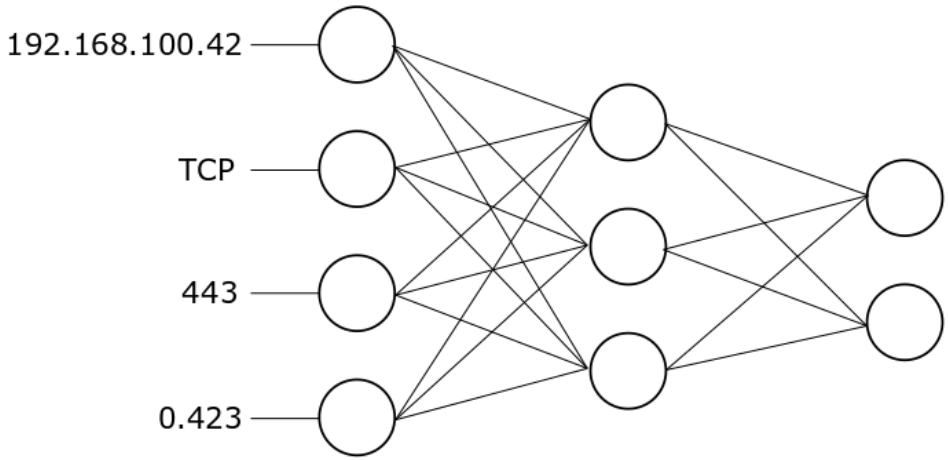
2) Approach
2.1) Generative Adversarial Networks

- Use GANs for creating flow-based network traffic



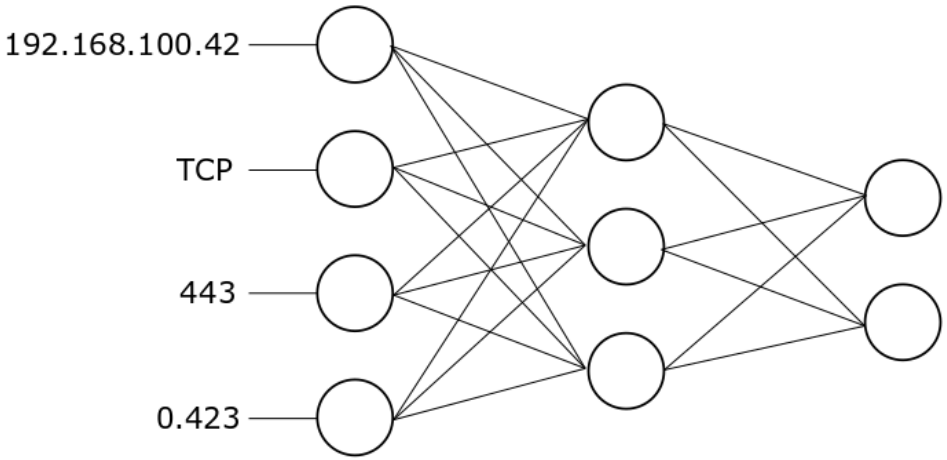
2) Approach
2.2) Challenges

GANs can only process continuous data but flow-based network traffic contains also categorical attributes like IP Addresses



2) Approach
2.2) Challenges

GANs can only process continuous data but flow-based network traffic contains also categorical attributes like IP Addresses



– Improved Wasserstein GANs* with the two time scale update rule from Heusel et al.+)

* Gulrajani, Ishaan ; Ahmed, Faruk ; Arjovsky, Martin ; Dumoulin, Vincent ; Courville, Aaron C.: Improved Training of Wasserstein GAN. In: Advances in Neural Information Processing Systems (NIPS), 2017, S. 57695779

+ Heusel, Martin ; Ramsauer, Hubert ; Unterthiner, Thomas ; Nessler, Bernhard ; Hochreiter, Sepp: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: Advances in Neural Information Processing Systems (NIPS), 2017, S. 66296640

2) Approach

2.3) Representation of flow-based network traffic

- Flows describe meta information about network connections between endpoint devices

#	Timestamp	Dur.	Proto	Source IP	Source Port	Dest. IP	Dest. Port	Bytes	Pkt.	TCP Flags
1	2019-03-07 13:41:41	0.000	UDP	192.168.220.16	48297	192.129.29.8	53	66	1	.....
2	2019-03-07 13:41:41	0.000	UDP	192.129.29.8	53	192.168.220.16	48297	66	1	.....
3	2019-03-07 13:41:42	4.199	TCP	192.168.220.16	53333	192.168.100.5	80	180	3	.A..SF
4	2019-03-07 13:41:42	4.324	TCP	192.168.100.5	80	192.168.220.16	53321	932	4	.A..SF

2) Approach

2.3) Representation of flow-based network traffic

- Flows describe meta information about network connections between endpoint devices

#	Timestamp	Dur.	Proto	Source IP	Source Port	Dest. IP	Dest. Port	Bytes	Pkt.	TCP Flags
1	2019-03-07 13:41:41	0.000	UDP	192.168.220.16	48297	192.129.29.8	53	66	1	.....
2	2019-03-07 13:41:41	0.000	UDP	192.129.29.8	53	192.168.220.16	48297	66	1	.....
3	2019-03-07 13:41:42	4.199	TCP	192.168.220.16	53333	192.168.100.5	80	180	3	.A..SF
4	2019-03-07 13:41:42	4.324	TCP	192.168.100.5	80	192.168.220.16	53321	932	4	.A..SF

- Three approaches
  1. Numeric Approach: Convert categorical attributes to numbers
  2. Binary Approach: Create binary attributes from categorical attributes
  3. Embedding Approach: Learn embeddings for categorical attributes
- Baseline
  1. Draw from the empirical probability distribution

2) Approach
2.3) Representation of flow-based network traffic

HOCHSCHULE COBURG
university of applied sciences

- WGAN Approaches
  1. Numeric approach – Converting categorical values to numbers

#	Timestamp	Dur.	Proto	Source IP	Source Port	Dest. IP	Dest. Port	Bytes	Pkt.	TCP Flags
1	2019-03-07 13:41:41	0.000	UDP	192.168.220.16	48297	192.129.29.8	53	66	1	.....
2	2019-03-07 13:41:41	0.000	UDP	192.129.29.8	53	192.168.220.16	48297	66	1	.....
3	2019-03-07 13:41:42	4.199	TCP	192.168.220.16	53333	192.168.100.5	80	180	3	.A..SF
4	2019-03-07 13:41:42	4.324	TCP	192.168.100.5	80	192.168.220.16	53321	932	4	.A..SF

transform *daytime* into seconds [0…86400) and normalize to the interval [0,1]

normalize *Duration* to interval [0,1]

create 3 binary attributes (isTCP, isUDP, isICMP)

normalize each octet of an *IP Address*

normalize *Ports* to interval [0,1]

$$\frac{80}{65536} = 0{,}001\ldots$$

normalize *Bytes* to interval [0,1]

normalize *Packets* to interval [0,1]

create 6 binary attributes (isSYN,…)

$$192.168.210.5 \rightarrow \frac{192}{255} = 0{,}75\ldots, \frac{168}{255} = 0{,}65\ldots \frac{210}{255} = 0{,}82\ldots \frac{5}{255} = 0{,}01\ldots$$

2) Approach
2.3) Representation of flow-based network traffic

- WGAN Approaches
  2. Binary approach – Extracting binary attributes from categorical values

#	Timestamp	Dur.	Proto	Source IP	Source Port	Dest. IP	Dest. Port	Bytes	Pkt.	TCP Flags
1	2019-03-07 13:41:41	0.000	UDP	192.168.220.16	48297	192.129.29.8	53	66	1	.....
2	2019-03-07 13:41:41	0.000	UDP	192.129.29.8	53	192.168.220.16	48297	66	1	.....
3	2019-03-07 13:41:42	4.199	TCP	192.168.220.16	53333	192.168.100.5	80	180	3	.A..SF
4	2019-03-07 13:41:42	4.324	TCP	192.168.100.5	80	192.168.220.16	53321	932	4	.A..SF

transform *daytime* into seconds [0…86400) and normalize to the interval [0,1]

normalize *Duration* to interval [0,1]

create 3 binary attributes (isTCP,…)

interpret *IP Addresses* as 32 Bit Integer

192.168.210.5 → 11000000 10101000 ….

interpret *Ports* as 16 Bit Integer

80 → 00000000 01010000

interpret *Bytes* as 32 Bit Integer

interpret *Packets* as 32 Bit Integer

create 6 binary attributes (isSYN,…)

2) Approach
2.3) Representation of flow-based network traffic

- WGAN Approaches
  3. Embedding approach – Learning embeddings using IP2Vec*

#	Timestamp	Dur.	Proto	Source IP	Source Port	Dest. IP	Dest. Port	Bytes	Pkt.	TCP Flags
1	2019-03-07 13:41:41	0.000	UDP	192.168.220.16	48297	192.129.29.8	53	66	1	.....
2	2019-03-07 13:41:41	0.000	UDP	192.129.29.8	53	192.168.220.16	48297	66	1	.....
3	2019-03-07 13:41:42	4.199	TCP	192.168.220.16	53333	192.168.100.5	80	180	3	.A..SF
4	2019-03-07 13:41:42	4.324	TCP	192.168.100.5	80	192.168.220.16	53321	932	4	.A..SF

transform *daytime* into seconds [0…86400) and normalize to the interval [0,1]

replace *Duration* with embedding

create 3 binary attributes (isTCP,…)

replace *IP Adresses* with embedding

replace *Ports* with embedding

replace *Byte* with embedding

replace *Packets* with embedding

create 6 binary attributes (isSYN,…)

* M. Ring, A. Dallmann, D. Landes, A. Hotho: **IP2Vec**: Learning Similarities between IP Adresses. In Proc. Workshop on Data Mining for Cyber Security (DMCS), International Conference on Data Mining (ICDM 2017), New Orleans, USA, 2017, 657-666.

3) Experiments and Results
3.1) Overview

HOCHSCHULE COBURG
university of applied sciences

- Experimental Evaluation
  - Create flow-based network traffic on the CIDDS-001* data set
  - Split the four weeks of the CIDDS-001 data set in two parts
    - week1 (reference data)
    - week2-4 (training data)

- Evaluation of generative models
  - Use different approaches for evaluation
    1. Temporal progression of the generated traffic
    2. Visualization of diversity and inter-attribute relationships
    3. Similarity calculations of value distributions
    4. **Intrinsic evaluation using domain knowledge checks**

* Ring, Markus ; Wunderlich, Sarah ; Grüdl, Dominik ; Landes, Dieter ; Hotho, Andreas: Flow-based benchmark data sets for intrusion detection. In: European Conference on Cyber Warfare and Security (ECCWS). ACPI, 2017, S. 361369

3) Experiments and Results

3.3) Intrinsic evaluation: Domain knowledge checks

accuracy in percent

	Baseline	Numeric	Binary	Embedding
Test 1	14.08	96.46	97.88	**99.77**
Test 2	81.26	0.61	98.90	**99.98**
Test 3	86.90	95.45	**99.97**	**99.97**
Test 4	15.08	7.14	**99.90**	99.84
Test 5	**100.0**	25.79	47.13	99.80
Test 6	0.07	0.00	40.19	**92.57**
Test 7	71.26	**100.0**	85.32	99.49

3) Experiments and Results

3.3) Intrinsic evaluation : Domain knowledge checks

accuracy in percent

	Baseline	Numeric	Binary	Embedding
Test 1	14.08	96.46	97.88	**99.77**
Test 2	81.26	0.61	98.90	**99.98**
Test 3	86.90	95.45	**99.97**	**99.97**
Test 4	15.08	7.14	**99.90**	99.84
Test 5	**100.0**	25.79	47.13	99.80
Test 6	0.07	0.00	40.19	**92.57**
Test 7	71.26	**100.0**	85.32	99.49

Test 1: If transport protocol is UDP, then the flow must not have any TCP Flags

3) Experiments and Results

3.3) Intrinsic evaluation : Domain knowledge checks

HOCHSCHULE COBURG
university of applied sciences

accuracy in percent

	Baseline	Numeric	Binary	Embedding
Test 1	14.08	96.46	97.88	**99.77**
Test 2	81.26	0.61	98.90	**99.98**
Test 3	86.90	95.45	**99.97**	**99.97**
Test 4	15.08	7.14	**99.90**	99.84
Test 5	**100.0**	25.79	47.13	99.80
Test 6	0.07	0.00	40.19	**92.57**
Test 7	71.26	**100.0**	85.32	99.49

Test 1: If transport protocol is UDP, then the flow must not have any TCP Flags

the baseline does not consider inter-attribute relationships

3) Experiments and Results

3.3) Intrinsic evaluation : Domain knowledge checks

HOCHSCHULE COBURG
university of applied sciences

accuracy in percent

	Baseline	Numeric	Binary	Embedding
Test 1	14.08	96.46	97.88	**99.77**
Test 2	81.26	0.61	98.90	**99.98**
Test 3	86.90	95.45	**99.97**	**99.97**
Test 4	15.08	7.14	**99.90**	99.84
Test 5	**100.0**	25.79	47.13	99.80
Test 6	0.07	0.00	40.19	**92.57**
Test 7	71.26	**100.0**	85.32	99.49

Test 2: At least one IP address of each flow must be internal
(starting with 192.168.XXX.XXX)

3) Experiments and Results

3.3) Intrinsic evaluation : Domain knowledge checks

HOCHSCHULE COBURG
university of applied sciences

accuracy in percent

	Baseline	Numeric	Binary	Embedding
Test 1	14.08	96.46	97.88	**99.77**
Test 2	81.26	0.61	98.90	**99.98**
Test 3	86.90	95.45	**99.97**	**99.97**
Test 4	15.08	7.14	**99.90**	99.84
Test 5	**100.0**	25.79	47.13	99.80
Test 6	0.07	0.00	40.19	**92.57**
Test 7	71.26	**100.0**	85.32	99.49

Test 2: At least one IP address of each flow must be internal
(starting with 192.168.XXX.XXX)

Numeric approach often generates IP addresses like
191.168.103.78
192.167.103.78

3) Experiments and Results

3.3) Intrinsic evaluation : Domain knowledge checks

accuracy in percent

	Baseline	Numeric	Binary	Embedding
Test 1	14.08	96.46	97.88	**99.77**
Test 2	81.26	0.61	98.90	**99.98**
Test 3	86.90	95.45	**99.97**	**99.97**
Test 4	15.08	7.14	**99.90**	99.84
Test 5	**100.0**	25.79	47.13	99.80
Test 6	0.07	0.00	40.19	**92.57**
Test 7	71.26	**100.0**	85.32	99.49

Test 6: If the flow represents a netbios message (destination port is 137 or 138),
the source IP address must be internal (192.168.XXX.XXX) and
the destination IP address must be an internal broadcast (192.168.XXX.255)

4) Summary

HOCHSCHULE COBURG
university of applied sciences

GANs are suitable for creating synthetic flow-based network traffic, if a good representation is chosen

4) Summary

HOCHSCHULE COBURG
university of applied sciences

GANs are suitable for creating synthetic flow-based network traffic, if a good representation is chosen

Advantages

- Enrich existing data sets with synthetic data
- Trained models may be passed on
- Can be trained on different data sets to combine them

4) Summary

HOCHSCHULE COBURG
university of applied sciences

GANs are suitable for creating synthetic flow-based network traffic, if a good representation is chosen

Advantages

- Enrich existing data sets with synthetic data
- Trained models may be passed on
- Can be trained on different data sets to combine them

Next Steps

- Handle sequences of flows

Thanks for your Attention!

Questions?

- GANs are able to create realistic flow-based network traffic
- How to create sequences of flows?
- Further information about network-based data

  https://www.hs-coburg.de/cidds (own data sets)

  http://www.dmir.uni-wuerzburg.de/datasets/nids-ds (overview of data sets)

E-Mail: markus.ring@hs-coburg.de