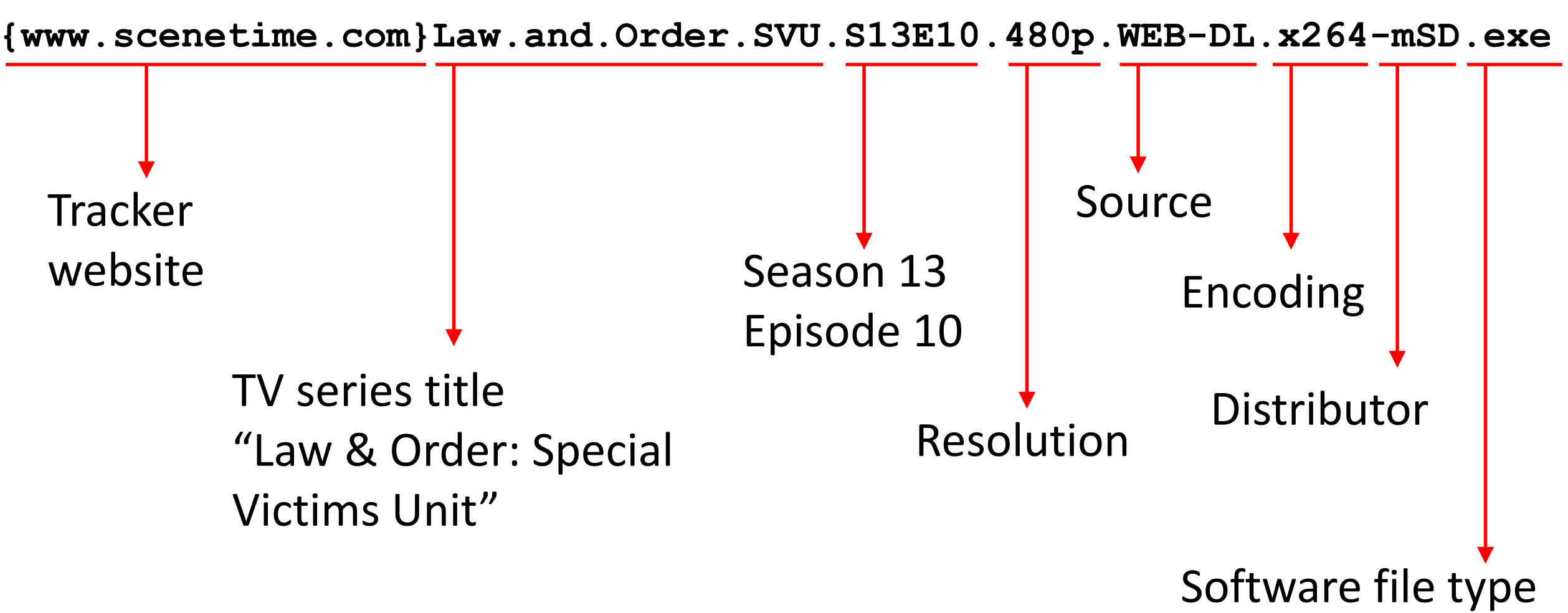
Bayesian Torrent Classification by File Name and Size Only

Eugene Dementiev, Norman Fenton

INTRODUCTION

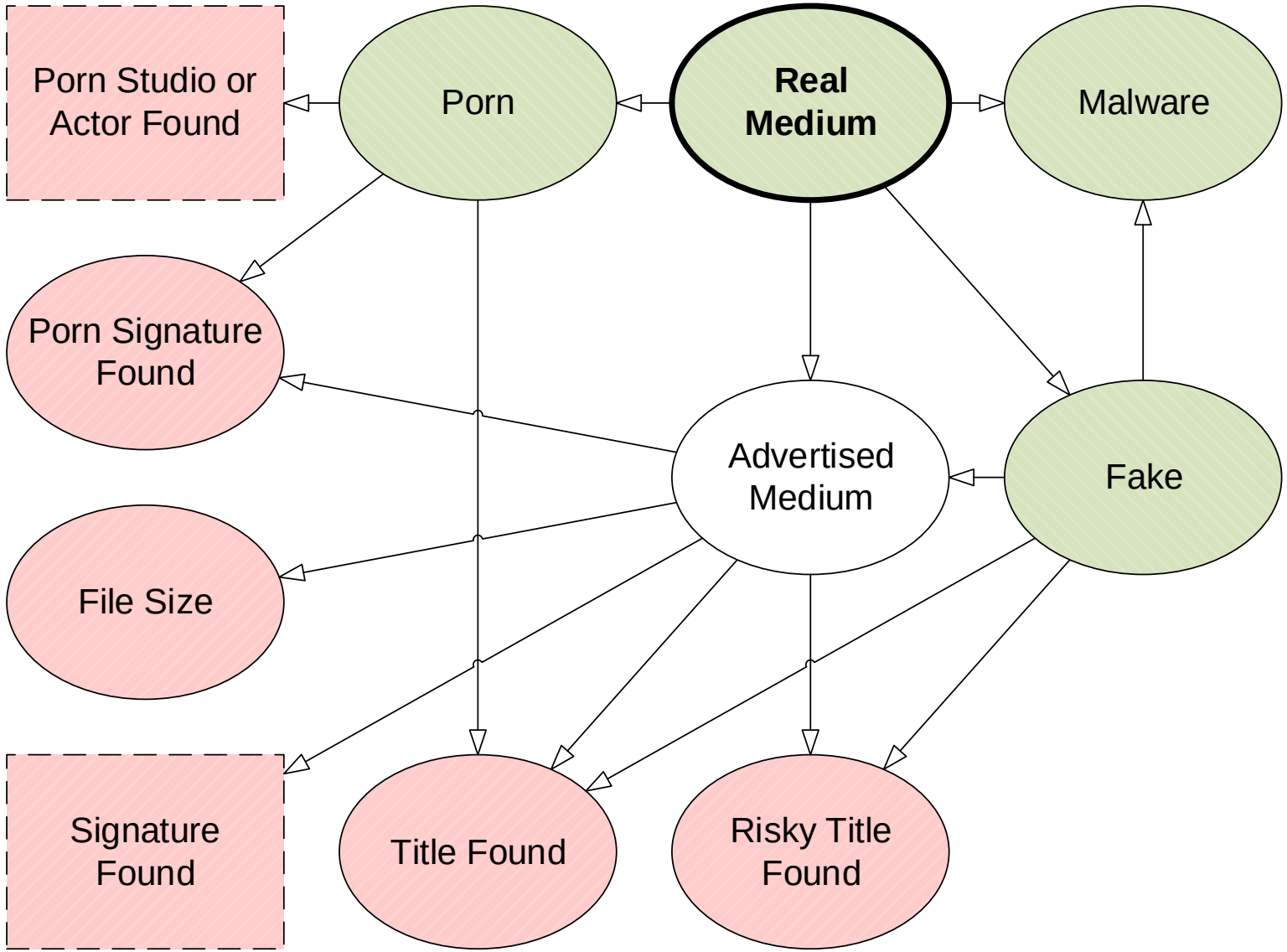
- Up to 30% of global traffic caused by torrents
- Claims of piracy and lost revenue motivate legislation such as PIPA/SOPA
- Identification and classification lacks transparency
- Potential to recognize malware threat even before download

TORRENT AND EVIDENCE EXAMPLE



BAYESIAN NETWORK

- Predict: medium, porn, fake, malware
- Intent: sharing vs deception
- Advertised vs real medium
- Dedicated evidence node for each class



Eugene Dementiev, Norman Fenton

BAYES KNOWLEDGE
EPSRC Engineering and Physical Sciences Research Council
Queen Mary University of London
agena
movie labs

EMPIRICAL EVALUATION 1

		Error		Accuracy	
Set	**Agent**	**Brier Score**	**Mean Absolute Error**	**Recall**	**Precision**
DS120	Humans	0.30 ±0.08	0.38 ±0.05	0.80	0.80
	MovieLabs	0.64 ±0.13	0.43 ±0.08	0.56	0.56
	Toran T	0.46 ±0.10	0.40 ±0.06	0.66	0.66
DSBS	MovieLabs	0.384 ±0.006	0.298 ±0.004	0.720	0.720
	Toran	0.269 ±0.005	0.231 ±0.003	0.811	0.811
	Toran T	0.268 ±0.005	0.217 ±0.003	0.816	0.816

Results comparison with 95% confidence intervals

Toran is our system with title matching turned off and Toran T – with title matching turned on

EMPIRICAL EVALUATION 2

		Error	Accuracy	
	Agent	**Mean Absolute Error**	**Recall**	**Precision**
Porn	Humans	0.13 ±0.03	0.94	0.97
	MVL	0.21 ±0.07	0.31	1.00
	Toran T	0.19 ±0.04	0.61	0.92
Fakes	Humans	0.38 ±0.08	0.30	0.64
	Toran T	0.29 ±0.08	0.57	0.84
Malware	Humans	0.17 ±0.07	0.26	1.00
	Toran T	0.18 ±0.07	0.30	0.88

Results comparison with 95% confidence intervals

Toran is our system with title matching turned off and Toran T – with title matching turned on

PGM'16

Thank you!

Eugene Dementiev, Norman Fenton

BAYES KNOWLEDGE
EPSRC Engineering and Physical Sciences Research Council
Queen Mary University of London
agena
movie labs