PASYDA - Synthetic Datasets for Enhancing Online Child Protection from Grooming*

The PASYDA project aimed to address the challenge of generating synthetic datasets capturing online grooming. With the rise of end-to-end encryption, accessing raw message data has become increasingly challenging. Therefore, the focus of the PASYDA project is to generate synthetic datasets of metadata associated with the messages being exchanged. Metadata is accessible and relevant for real-world applications. The primary objective is to provide researchers with a valuable resource for refining and validating their detection models for online grooming.

To accomplish this goal, we sought to create an agent-based simulation to simulate interactions within the victim's social network. For generating the synthetic datasets two sources were used as data sources of grooming cases: Lottie which is a training simulation[1], and Perverted Justice[2] which is a collection of cases where adults pose as children to out potential groomers.

A method for generating a synthetic social network was developed by analysing the SNAP Social Circles[3] dataset for extracting distributions to inform the structure of the synthetic network.

Upon establishing a grooming scenario and generating the social network, we developed an agent-based simulation using NetLogo[4]. This simulation populated the network with agents representing individuals and simulated their social media engagements, thereby generating a stream of interactions. These simulated messages were recorded to form our synthetic dataset.

The outcomes of our project are threefold:

1. We present 11 synthetic datasets encapsulating metadata from social media interactions centred around instances of child grooming. These datasets serve as resources for researchers seeking to evaluate and enhance their detection algorithms.

2. We offer 11 examples of simulation files, defining the parameters utilized in the generation of the synthetic datasets. These files serve as blueprints for researchers who may want to replicate or modify our methodology.

3. We provide supporting code for generating synthetic networks and conducting comprehensive analyses of the datasets. This code facilitates the analysis of the generated datasets. Moreover, it allows researchers to generate further data and analyse the results of the simulation to investigate features of the conversation data and compare the distributions of the generated data to that of the grooming case.

---

*Rogério de Lemos, Virginia Franqueira, Tracee Green, Marek Grzes and Robin Ayling; project sponsored by REPHRAIN (`https://www.rephrain.ac.uk/`)

[1]Reeves, J., Shemmings, D. and Blake, E. (2014), 'Looking out for Lottie: Child Sexual Exploitation'. University of Kent. Available at: `http://www.kent.ac.uk/sspssr/ccp/game/Lottieindex.html`.

[2]Perverted Justice, `http://www.perverted-justice.com/`

[3]McAuley, J. and Leskovec, J.. Learning to Discover Social Circles in Ego Networks. NIPS, 2012. `http://i.stanford.edu/~julian/pdfs/nips2012.pdf`

[4]NetLogo, `https://ccl.northwestern.edu/netlogo/`