# Generative Deep Learning for Solutions to Data Deconflation Problems in Information and Operational Technology Networks

Roger A. Hallman[1,3], John M. San Miguel[2], Arron Lu[2], Alejandro Monje[2], Mohammad R. Alam[4] and George Cybenko[3]

[1]*C.A.T Labs, San Diego, California, U.S.A.*

[2]*Naval Information Warfare Center Pacific, San Diego, California, U.S.A.*

[3]*Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire, U.S.A.*

[4]*The M.I.T.R.E. Corporation, San Diego, California, U.S.A.*

Keywords: Data Deconflation, Source Separation, Generative Adversarial Networks (GANs), Transformers, Double-NATed Network Traffic, Network Situational Awareness.

Abstract: Source separation problems are a long-standing and well-studied challenge in signal processing and information sciences. The "Cocktail Party Phenomenon" and other classical source separation problems are vector representable and additive, and thus solvable by well-established linear algebra techniques. However, the proliferation and adoption of Internet-connected devices (e.g., IoT, distributed sensor networks, etc.) have led to a "Cambrian explosion" of data that is available for processing. Much of this data is not readily available for processing because it includes data objects that are categorical or non-additive superpositions (i.e., data not confined to signals). The Data Deconflation Problem refers to the challenge of identifying and separating the individual constituent elements of these complex data objects. Real-world data deconflation scenarios include pattern-of-life tracking (e.g., identifying recreational activities in conjunction with a business trip), multi-target tracking (e.g., occlusions and track assignment challenges), and network situational awareness (e.g., monitoring NATed network traffic, detecting and identifying shadow IT, network steganalysis).
This paper details our approach, utilizing Generative Adversarial Networks (GANs) and attention-based Transformers, to solving the data deconflation problem, as well as our experimental application to network situational awareness tasks. We cover traditional source separation solutions and expound upon why these solutions are inadequate for network monitoring tasks. Background information on GANs and transformers is presented before a description of our architecture and initial experimentation which serves as a proof-of-concept. We then describe experimentation applying our methodology to network monitoring tasks, in particular separating activities and shadow IT devices within double-NATed network traffic. We discuss our results and our methodology's applicability to other network monitoring tasks, such as network steganalysis and covert channel detection.

## 1 INTRODUCTION AND MOTIVATION

The ever-increasing adoption of distributed sensor networks, Internet of Things (IoT), networked infrastructure, as well as mobile and wearable devices has brought about a "Cambrian Explosion" of data that is available for processing. It is unlikely that these torrents of data will be ready–or even useful–for processing and computation upon arrival at data centers. For instance, wearable medical devices may have readings corrupted by patient movement, data from sensor networks may represent co-located individuals, or IP addresses made 'private' by a router's network address translation (NAT)–a solution to the depletion of IPv4 addresses–are obfuscated and it is difficult to identify individual NATed machines. There is a particularly interesting variation of the last example where two routers are placed sequentially in a network's architecture called a *double-NAT* (Karimzadeh et al., 2017). Double-NATing may be intentionally designed into a a network architecture, but it often added in later, leading to a condition called *shadow IT* which is difficult for network administrators to deal



Hallman, R., Miguel, J., Lu, A., Monje, A., Alam, M. and Cybenko, G.
Generative Deep Learning for Solutions to Data Deconflation Problems in Information and Operational Technology Networks.
DOI: 10.5220/0011996700003482
In *Proceedings of the 8th International Conference on Internet of Things, Big Data and Security (IoTBDS 2023)*, pages 231-235
ISBN: 978-989-758-643-9; ISSN: 2184-4976
Copyright © 2023 by SCITEPRESS – Science and Technology Publications, Lda. Under CC license (CC BY-NC-ND 4.0)

with as it obscures their visibility of networked devices. Shadow IT is defined as any solution (software, hardware, optimization, etc.) on a network that has not been approved by network administrators (Silic and Back, 2014). The challenge of discovering and classifying shadow IT devices is critical for enterprise network security.

To meet this challenge, we are utilizing the Machine Learning for Data Deconflation (ML4DD) approach that was introduced in (Hallman. and Cybenko., 2021). ML4DD is utilizing recent advances in deep learning to create novel solutions to the data deconflation problem, an updated take on classical source separation problems. Whereas classical source separation problems could be solved using well-established linear algebra techniques, however there are many real-world cases of conflated data that are mixed with important components that are not vector representable. Our approach attempts to address this challenge by utilizing Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Tomczak, 2022) that can observe streams of conflated data and create candidate processes that are likely to be responsible for creating the observed data stream. We have conducted initial experimentation and present our results as a proof-of-concept. We then describe ongoing experimentation building on our initial results to monitor traffic and identify individual devices contributing to double-NATed network.

The remainder of this paper is organized as follows: Background information and related work are presented in Section 2. The ML4DD concept and architecture, as well as initial proof-of-concept results are discussed in Section 3. We discuss the deconflation of IT/OT network traffic in Section 4 Finally, concluding remarks are presented in Section 5, along with directions for follow-on work.

## 2 BACKGROUND AND RELATED WORK

Data can be conflated in multiple dimensions (e.g., time, space, semantics, etc.) and there are many common manifestations. Consider the scenario of an employee using a company-owned computer on an enterprise network. This is an example of semantic data conflation, common to pattern-of-life analysis, where the employee will very likely be simultaneously running business applications (e.g., working in a spreadsheet, reading or writing business documentation) as well as a web browser with tabs open to recreational services (e.g., personal email, music or video streaming services, news aggregation sites, social media sites).

Many classical deconflation problems are solved using well-established linear algebra techniques–i.e., blind source separation (BSS) (Kofidis, 2016). In BSS, we seek to solve for a mixture

$$\mathbf{u}(n) = \mathcal{F}\left(\mathbf{a}(n), \mathbf{v}(n), n\right)$$

mixes $N$ source signals

$$\mathbf{a}(n) = \left[a_1(n), a_2(n), ..., a_N(n)\right]^T,$$

and $K$ noise signals

$$\mathbf{v}(n) = \left[v_1(n), v_2(n), ..., v_K(n)\right]^T,$$

by a mixing system $\mathcal{F}(\cdot,\cdot,\cdot)$, which yields

$$\mathbf{u}(n) = \left[u_1(n), u_2(n), ..., u_{N\times K}(n)\right]^T.$$

BSS has been successfully applied to signal processing applications across multiple modalities (e.g., the cocktail party phenomena, multimedia steganalysis, etc.).

Process Query Systems (PQS) (Cybenko and Berk, 2007), the current state-of-the-art deconflation solution, are well-suited to applications in networked environments. PQS work for discovering processes with discrete states, observable events, and dynamics. Multiple hypotheses are built about the processes behind observed events by taking inputs from arbitrary network nodes, ideally matching hypotheses to known processes. There are many PQS implementations used for covert channel detection (Giani et al., 2005), as well as other computer and network security applications (Berk et al., 2003; Berk and Fox, 2005). Despite these successful implementations, PQS require significant background information (e.g., a priori models, process heuristics, etc.) to be effective.

## 3 MACHINE LEARNING FOR DATA DECONFLATION: OUR APPROACH

The ML4DD approach to data deconflation was introduced in (Hallman. and Cybenko., 2021), incorporating recent advances in deep learning to move towards a more generalized solution to the data deconflation problem. Our approach takes the same fundamental assumptions that underlie PQS, namely that observed events in an environment are representative of underlying and interleaved data and processes. We take sequences of observed events as inputs and use a transformer-enhanced generative adversarial network (GAN) architecture generates candidates subsequences which represent possible underlying processes and data objects.

### 3.1 Early Results and Ongoing Experimentation

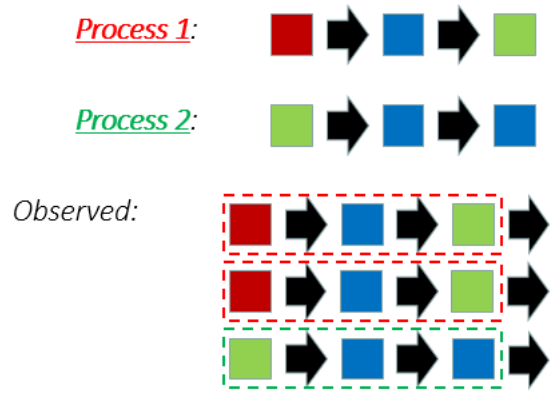


Figure 1: An observed sequence of two simple, interleaved processes for initial experimentation.

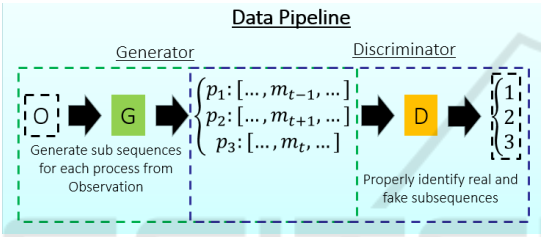


Figure 2: The ML4DD data pipeline for generating candidate process sequences.

We set up a simple illustrative example to demonstrate the ML4DD proof-of-concept. We begin with two repeating simple processes with observable states (Figure 1 top):

- Process 1 proceeds through its states as follows: $RED \rightarrow BLUE \rightarrow GREEN$;
- Process 2 proceeds through its states as follows: $GREEN \rightarrow BLUE \rightarrow BLUE$.

The processes are mixed according to some unknown probability to produce a sequence of observed events (Figure 1 top). The ML4DD GAN ingests this sequence of observed events and generates Process 1 and Process 2 after a sufficient number of rounds (Figure 2).

## 4 DECONFLATION OF NETWORK TRAFFIC

Following the promising results of our initial experimentation, we are utilizing ML4DD to give improved situational awareness for network monitoring tasks. In particular, we are interested in improved detection and identification of shadow IT assets. We are conducting experimentation on a dataset (Farhat et al., 2020) of double-NATed network traffic, with the end goal of being able to identify devices that are obscured behind a second router. This dataset replicates a shadow IT scenario in an enterprise network and records the network traffic from different devices: a PC, an ios device, and two different Android devices. There are 294 test sessions taken over one week, with each session consisting of seven tests recording network traffic over one-minute testing intervals.

We model network traffic as non-determinate finite state automata (Figure 3). TCP traffic can only take a finite number of states, so this is an ideal way to intake network traffic in a way that the ML4DD architecture can process. Our automata model is capable of simultaneously ingesting and modeling multiple TCP streams, a prerequisite for deployments in real world networks. Importantly, in the shadow IT scenario, multiple TCP streams will emanate from a single IP address (i.e., the illicit router that was installed on the network).

We are in the process of conducting experimentation where data from the aforementioned dataset of Double-NATed network traffic. Our GAN is trained initially with automata models of TCP streams of individual devices, before receiving models of double-NATed network traffic. Once this experimentation is completed, we will extend our methodology to network steganalysis applications (i.e., detecting covert channels), as well as designing an implementation of our data deconflation architecture that can be used to monitor traffic in operational (e.g., SCADA) networks. Critical infrastructures are reliant on these networks; however, they are notoriously fragile, and therefore not well-suited to the security solutions that are available for IT networks. We anticipate that ML4DD will prove to be a capable tool for analyzing OT network traffic, while minimally impacting operational performance, and proactively detecting potentially malicious traffic.

## 5 CONCLUSION AND FUTURE WORK

We are leveraging recent advances in deep learning, particularly GANs, to solve deconflation/source separation problems that are inherent to networked systems (though we hope that our approach will eventually prove to be a general solution to all classes of deconflation problems). These deconflation problems are becoming increasingly important as “smart” Internet-connected technologies and distributed sen-

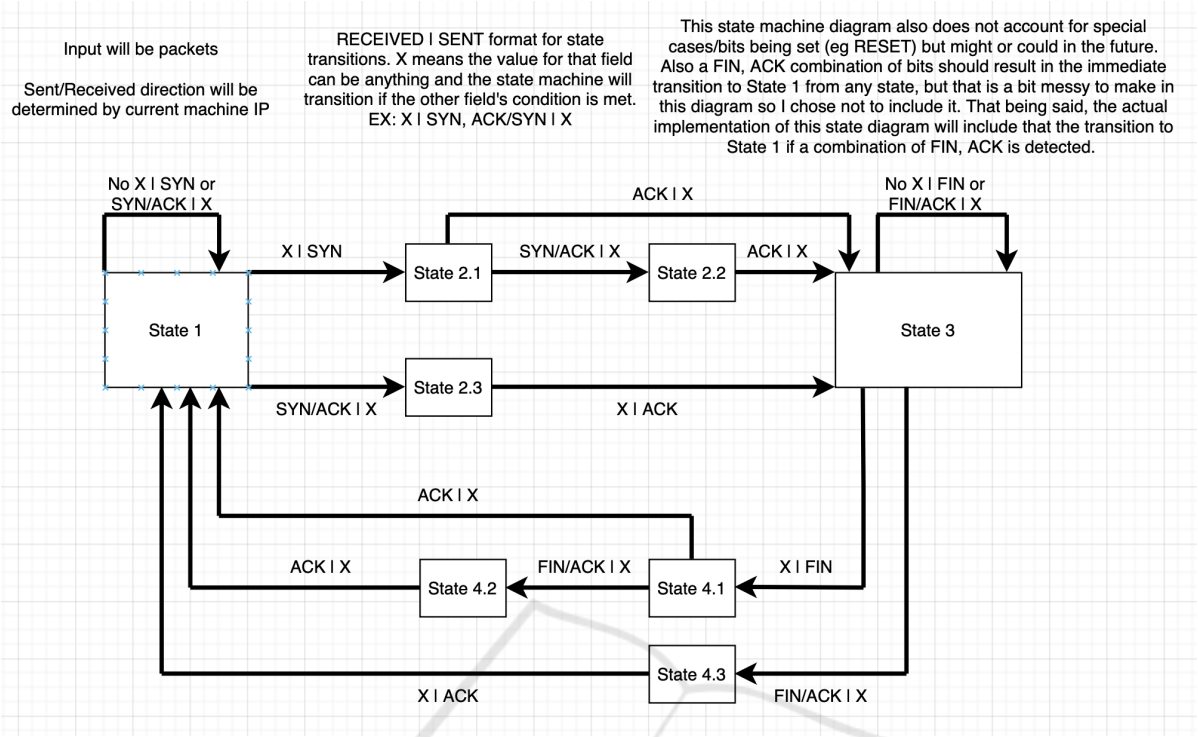


Figure 3: Automata model for TCP traffic.

sor networks are incorporated into real-world systems. This paper describes our initial results at deconflating simple processes, as well as our ongoing work with network monitoring use cases. In particular, we are pursuing the challenges of detecting and classifying shadow IT and covert channels on enterprise networks, as well as challenges that are unique to the analysis and defense of OT networks.

## ACKNOWLEDGEMENTS

Roger A. Hallman's contribution to this work occurred while employed by the Naval Information Warfare Center Pacific, during which time he was partially supported by the United States Department of Defense SMART Scholarship for Service Program, funded by USD/R&E (The Under Secretary of Defense-Research and Engineering), National Defense Education Program (NDEP) / BA-1, Basic Research.

## REFERENCES

Berk, V., Chung, W., Crespi, V., Cybenko, G., Gray, R., Hernando, D., Jiang, G., Li, H., and Sheng, Y. (2003). Process query systems for surveillance and awareness. In *In Proc. System. Cyber. Infor.(SCI2003*. Citeseer.

Berk, V. and Fox, N. (2005). Process query systems for network security monitoring. In *Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense IV*, volume 5778, pages 520–530. SPIE.

Cybenko, G. and Berk, V. H. (2007). Process query systems. *Computer*, 40(1):62–70.

Farhat, S., Elhajj, I. H., and Kayssi, A. (2020). Nat network traffic dataset. https://dx.doi.org/10.21227/zxdq-hg05.

Giani, A., Berk, V., Cybenko, G., and Hanover, N. (2005). Covert channel detection using process query systems. *proceedings of: FLoCon*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Hallman., R. and Cybenko., G. (2021). The data deconflation problem: Moving from classical to emerging solutions. In *Proceedings of the 6th International Conference on Internet of Things, Big Data and Security - AI4EIoTs,*, pages 375–380. INSTICC, SciTePress.

Karimzadeh, M., Valtulina, L., Pras, A., Liebsch, M., Taleb, T., van den Berg, H., and Schmidt, R. d. O. (2017). Double-nat based mobility management for future lte networks. In *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6. IEEE.

Kofidis, E. (2016). Blind source separation: Fundamentals and recent advances (a tutorial overview presented at sbrt-2001). *arXiv preprint arXiv:1603.03089*.

Silic, M. and Back, A. (2014). Shadow it–a view from behind the curtain. *Computers & Security*, 45:274–283.

Tomczak, J. M. (2022). *Deep Generative Modeling*. Springer Cham.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS