ML Enabled Cyber Deception

Kristen Moore

Senior Research Scientist
Distributed Systems Security

DecaaS: Deception as a Service

CYBER SECURITY
COOPERATIVE
RESEARCH
CENTRE

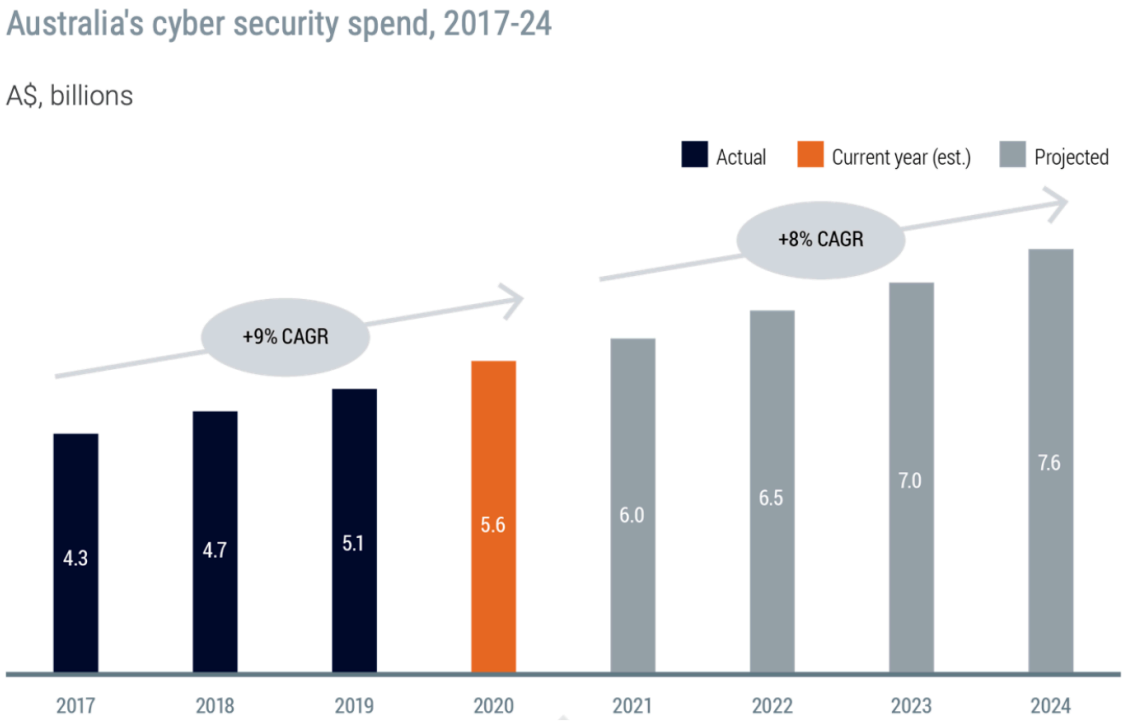
Research Partners –

Industry Partner –

Why Cyber Deception?

Australia spent **AU$5.6 billion** on cyber security in **2020**

Yet breaches continue…



Source: Australia's Cyber Security Sector Competitiveness Plan 2020, Australian Cyber Security Growth Network

Why Cyber Deception?

The average time to identify and contain a data breach is 277 days

The average cost of a data breach is US $4.35M

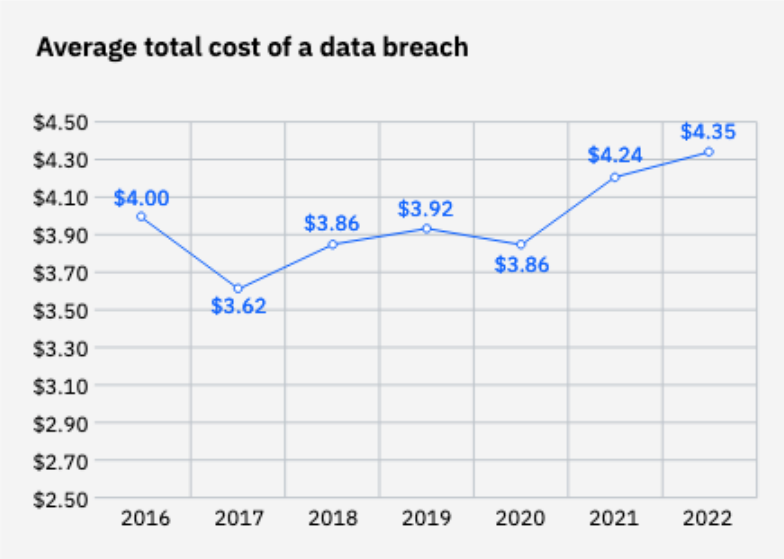


Figure 1: Measured in USD millions

Source: IBM/Ponemon **Cost of a Data Breach Report** 2022.

Deception for Security

Deception complements existing security technologies

Lets the defender regain the advantage

Honeypots

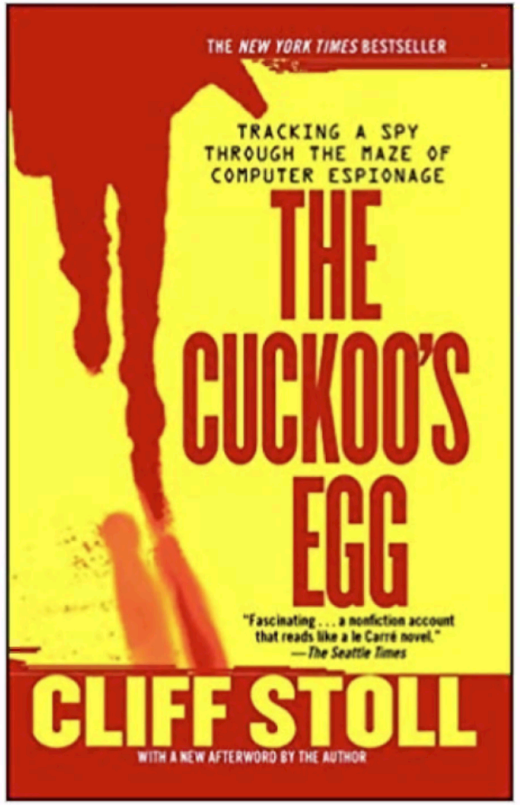
Cyber Deception typically means honeypots

- Put fake artefacts/traffic on the network
- Legitimate users have no reason to interact with
- Any interaction is suspicious → Breach discovery

Other Advantages

- Discover adversary intent
- Tactics, Tools and Procedures
- Delay and frustrate

A Cyber Deception success story:
The Cuckoo's Egg by Cliff Stoll



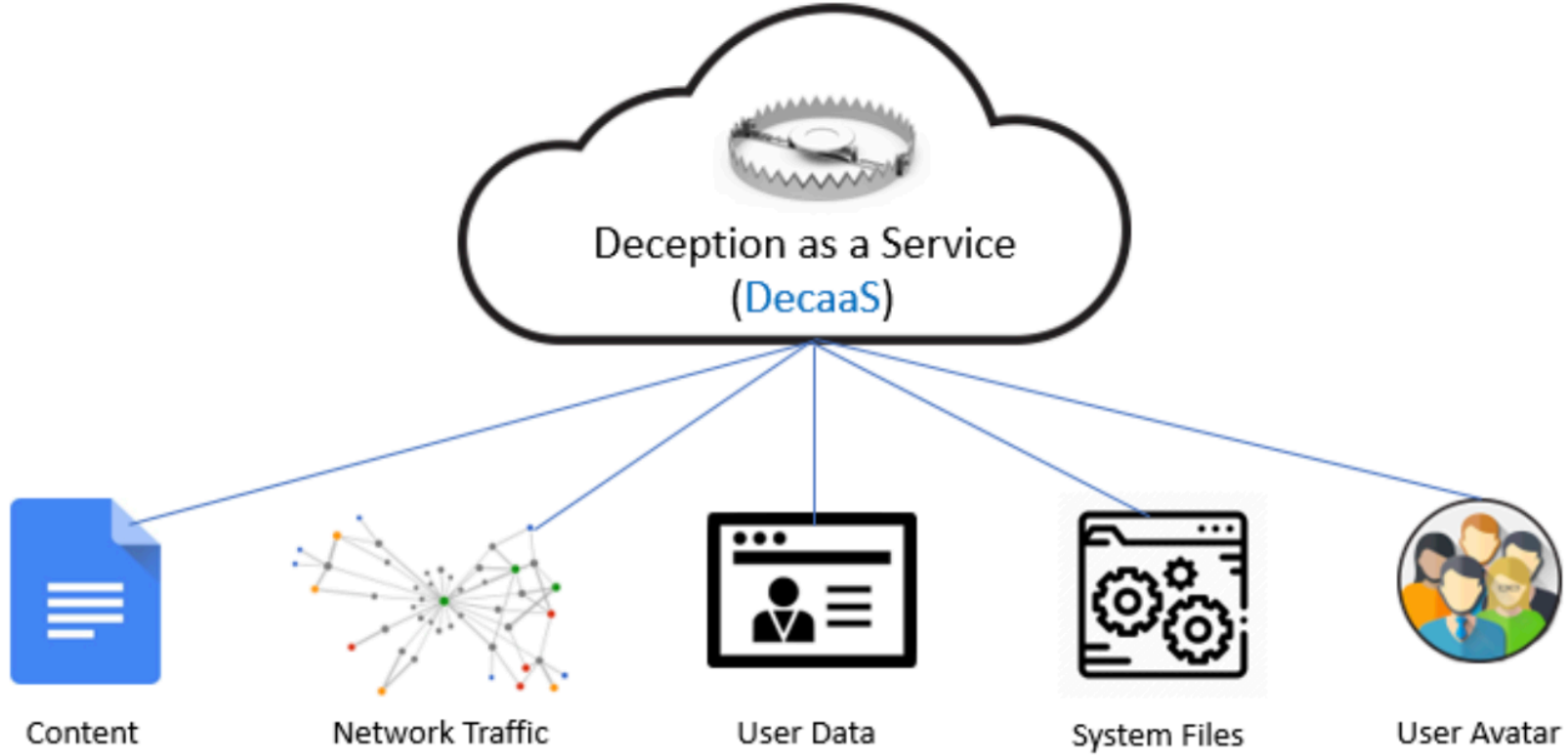
Technical Challenges

- **Realism** drives deep interaction
- **Automation** needed for scale

  → a Machine Learning problem…

CYBER SECURITY
COOPERATIVE
RESEARCH
CENTRE

DecaaS projects

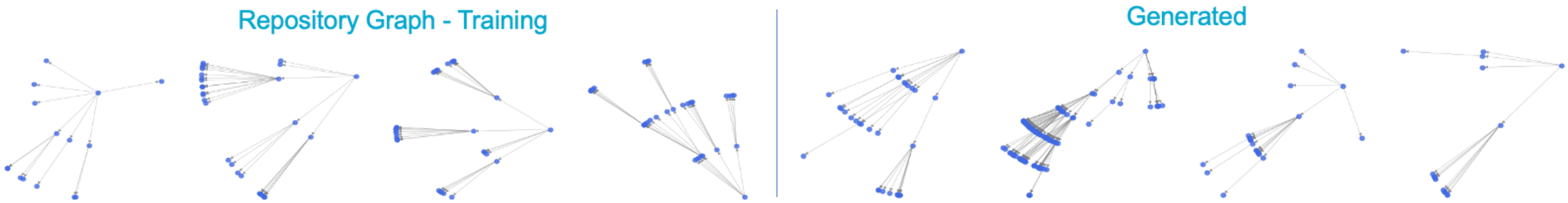


Key Machine Learning tools

- Language models (GPT, BERT, …)
- Graphs
- Temporal Point Processes (TPPs)

1. HoneyCode: Fake Repositories

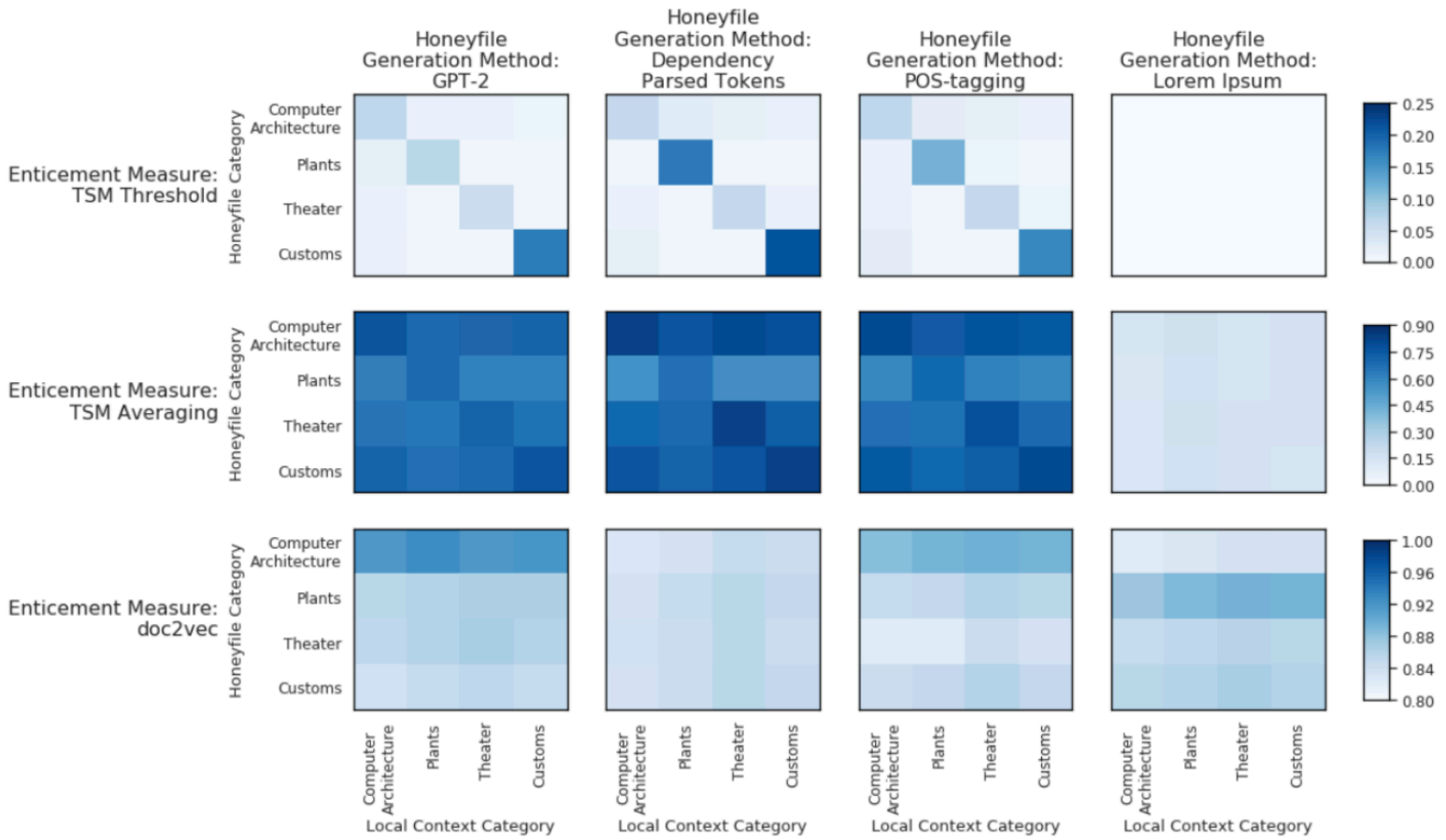
- Fake file trees, file names and code
- Trees generated using modified Graph Recurrent Attention Network (GRAN)
- Filenames and code from character RNN



See: D. Nguyen, D. Liebowitz, S. Nepal, and S. Kanhere, *Honeycode: Automating deceptive software repositories with deep generative models*, in Proceedings of the 54th Hawaii International Conference on System Sciences, 2021

2. Deception Metrics

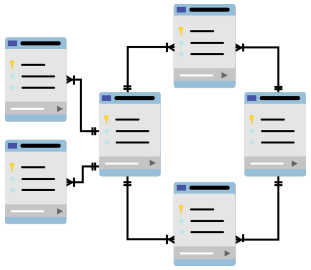
- Enticement: drawing adversary to the honeypot
- Compare honeyfile text to real files
- Topic modelling, semantic matching



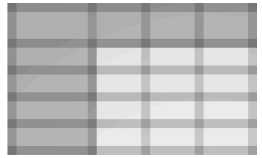
See: R. Timmer, D. Liebowitz, S. Nepal, and S. Kanhere, TSM: Measuring the Enticement of Honeyfiles with Natural Language Processing, accepted to: Proceedings of the 55th Hawaii International Conference on System Sciences, 2022

And more…

- Database Generation



- CSV content generation

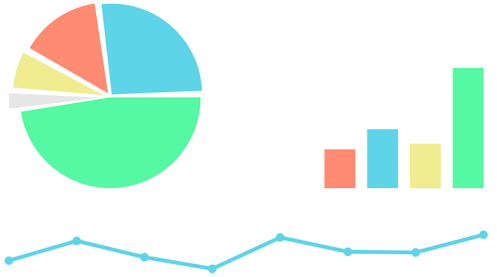


- Image and Logo Generation

- Wiki generation

- WiFi traffic generation

- Chart generation

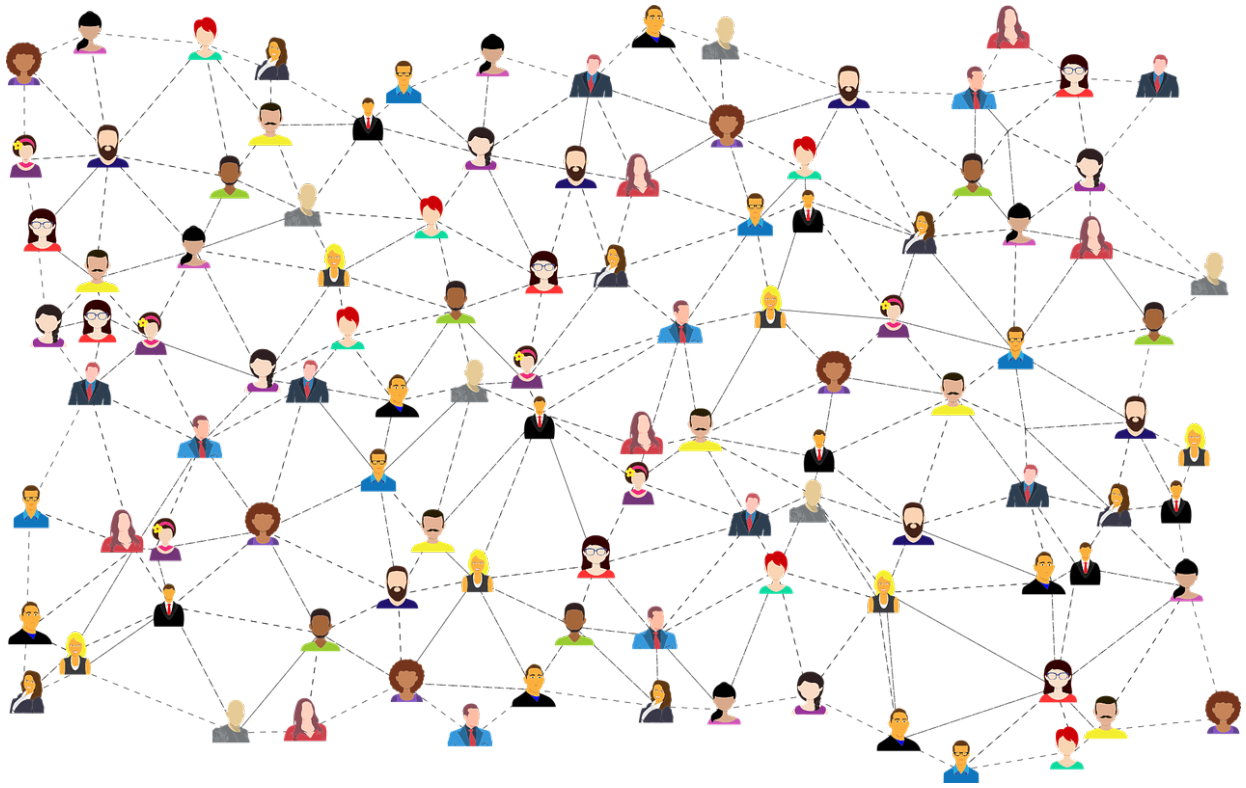


Simulating Networked Communications

Goal: Simulate communications on e-mail, Teams, Slack, Whatsapp, …

Approach: Combine:

- Temporal event models (including network topology)
- Language models



See: K. Moore, C. Christopher, D. Liebowitz, S. Nepal, and R. Selvey.
Modelling direct messaging networks with multiple recipients for cyber deception.
IEEE European Symposium on Security and Privacy 2022.

E-mail server simulation

e-Mail Generation

Timestamp and Participants Generator

Date:	From:	To:
Mon 10.15am	Marlene Evans	Phil Brooks
Mon 10.25am	Sarah Jones	Jo Ng; Bill Gates
...	...	...
...	...	...

e-Mail Subject and Content Generator

Subject: Lunch tomorrow

Hi Phil,

Shall we go to Szechuan House?

Regards,
Marlene

1. Simulating timestamps and participants

Date:	From:	To:
Mon 10.15am	Marlene Evans	Phil Brooks
Mon 10.25am	Sarah Jones	Jo Ng; Bill Gates
...	...	...
...	...	...

Part 1: Temporal event modeling

Real world event sequences

- Earthquakes
- User behaviours in social networks
- Patient Flows in Hospitals

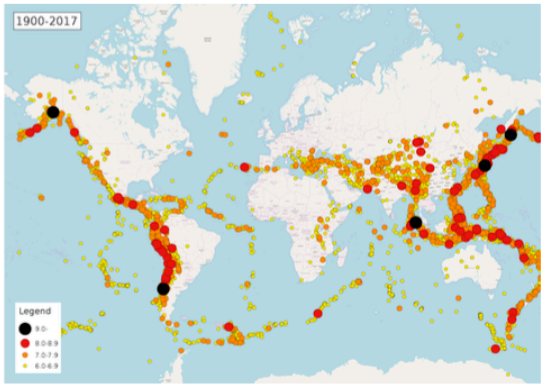


Figure 1: The locations and the intensities of the earthquakes from 1900 to 2017 [Ogata(1988)].

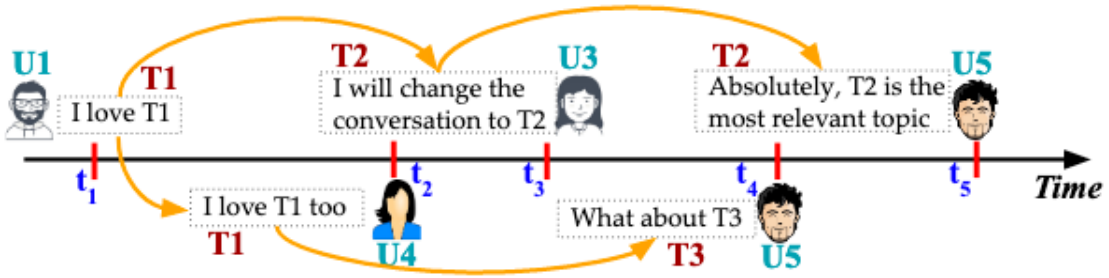


Figure 2: The times and topics of tweets of users on Twitter [J. Choudhari et al (2021)]

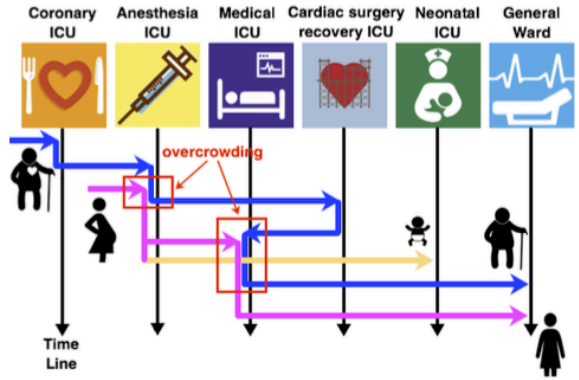
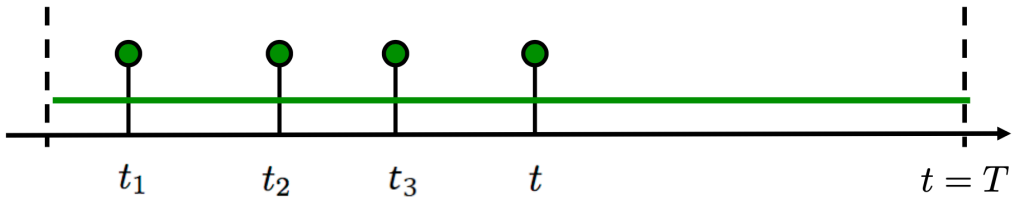


Figure 3: The transition behaviors of patients among different care units [Xu et al.(2016)a].

Temporal Point Processes

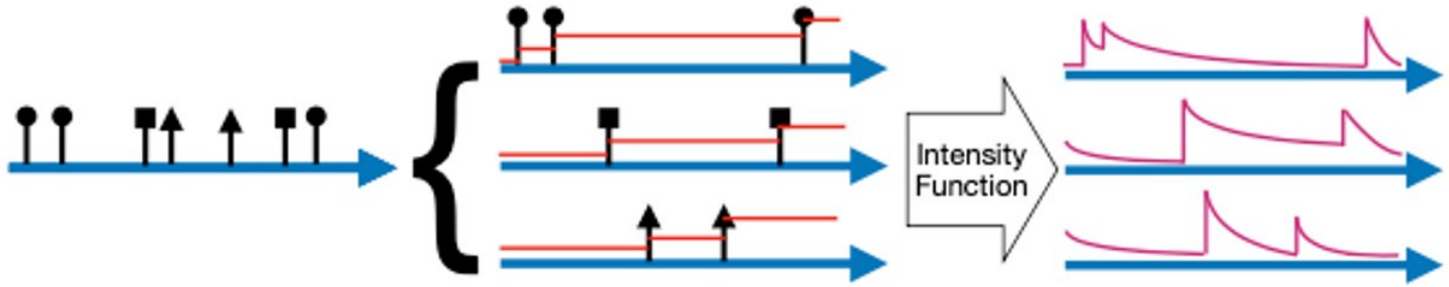
- Poisson process:



Constant rate intensity function

Learning with Temporal Point Processes. Rodriguez, M. G & Valera, I. *ICML 2018*

- Marked self-exciting TPPs:



Modelling and Applications for Temporal Point Processes. Xu, H. *KDD 2019*

- Contemporary approaches: neural networks (eg. RNNs or Transformers)

Desirable properties of TPP models

1. **Flexibility:** ability to approximate any probability density on $\mathbb{R}$ arbitrarily well. Eg. Multi-modal ones.
2. **Closed form likelihood:** if not closed form → have to approximate via Monte Carlo or numerical quadrature (slower and less accurate).
3. **Closed form sampling:** ie. draw samples analytically via inversion sampling. Alternative → Thinning Algorithm (slower and less accurate).

→ model the **conditional density** $p^*(t)$ using a **mixture model**
[Shchur et al. ICLR 2020]

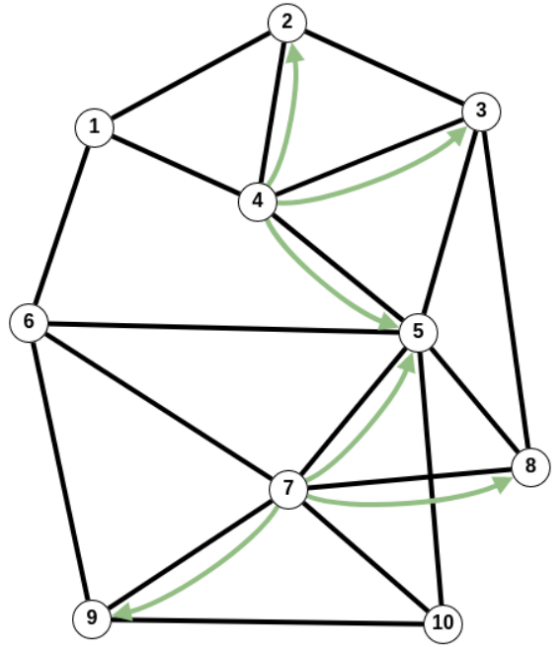
TPPs for network communication modeling

- **Multi-task ML problem with 3 tasks:**
  - event time prediction
  - sender prediction
  - recipient prediction
- **Event history:** embedded via RNN and shared across all 3 tasks

Date:	From:	To:
Mon 10.15am	Marlene Evans	Phil Brooks
Mon 10.25am	Sarah Jones	Jo Ng; Bill Gates
...	...	...
...	...	...

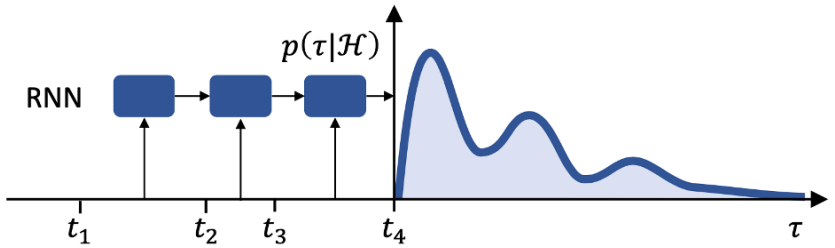
TPPs for network communication modeling

- **Multi-task ML problem with 3 tasks:**
  - event time prediction
  - sender prediction
  - recipient prediction
- **Event history:** embedded via RNN and shared across all 3 tasks



Date:	From:	To:
Mon 10.15am	Marlene Evans	Phil Brooks
Mon 10.25am	Sarah Jones	Jo Ng; Bill Gates
...	...	...
...	...	...

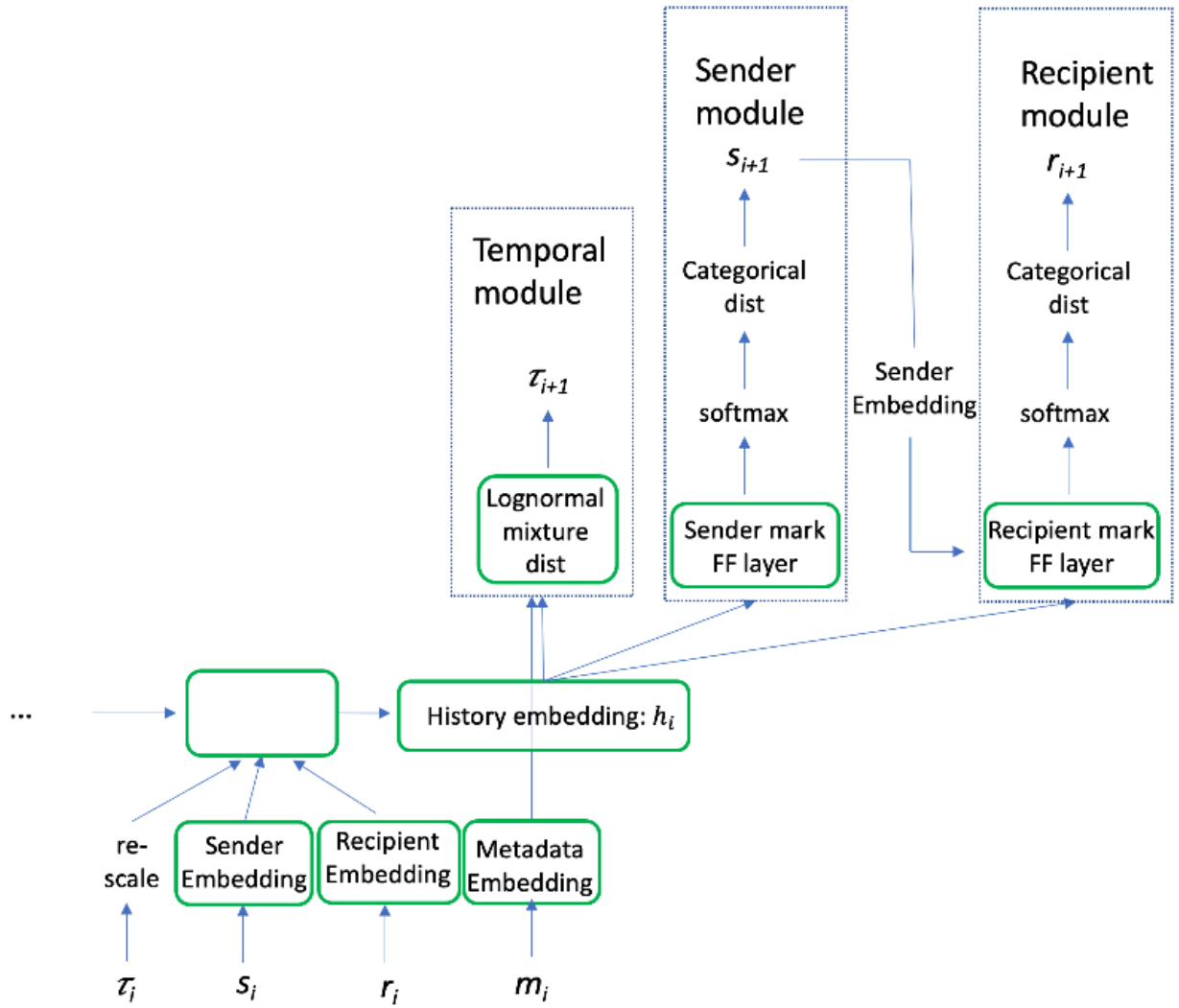
Time module:
Lognormal mixture model



Sender module:
Multi-class classification

Recipient module:
Multi-class classification
Conditioned on the e-Mail sender

LogNormMix-Net architecture

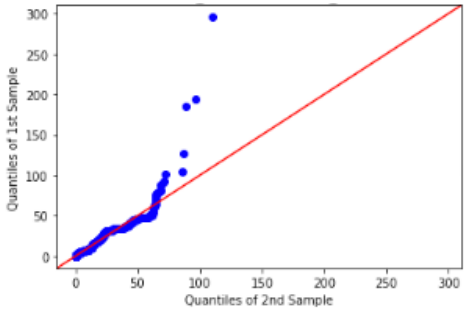


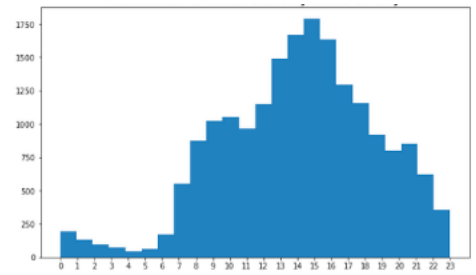
Evaluating the realism of generated content

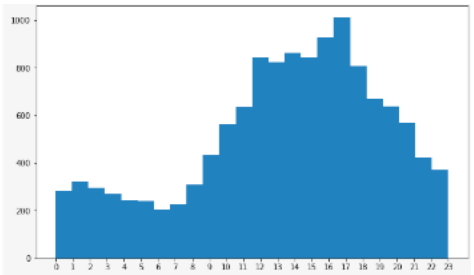
Cyber deception → simulated content should stand up to moderate scrutiny

Temporal realism:

- inter-arrival time distributions
- hour of day
- day of week





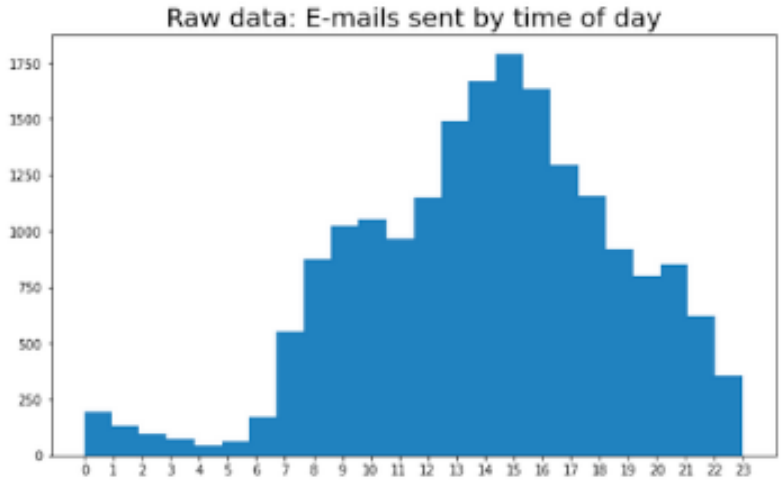


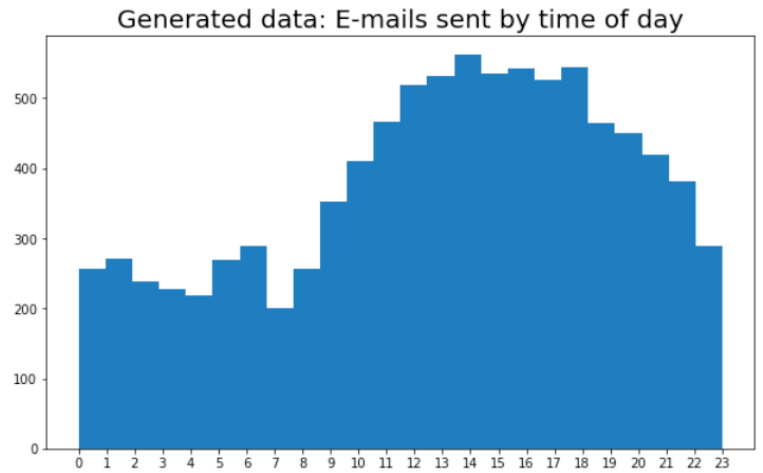
Training Vs Generated

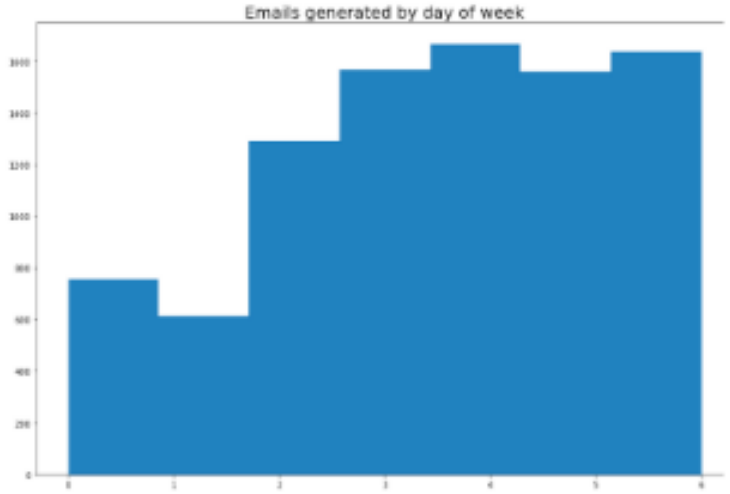
Realism for participants:

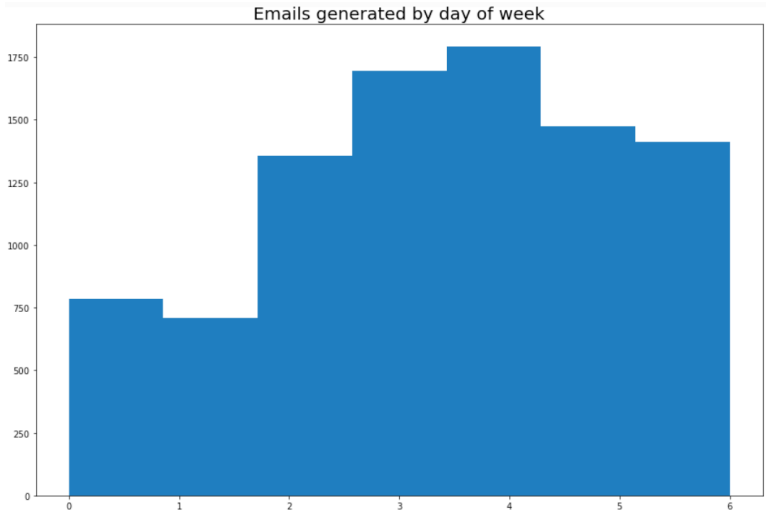
- Proportion of sent e-mails per node/user
- Proportion of received e-mails per recipient group

Seasonality preservation

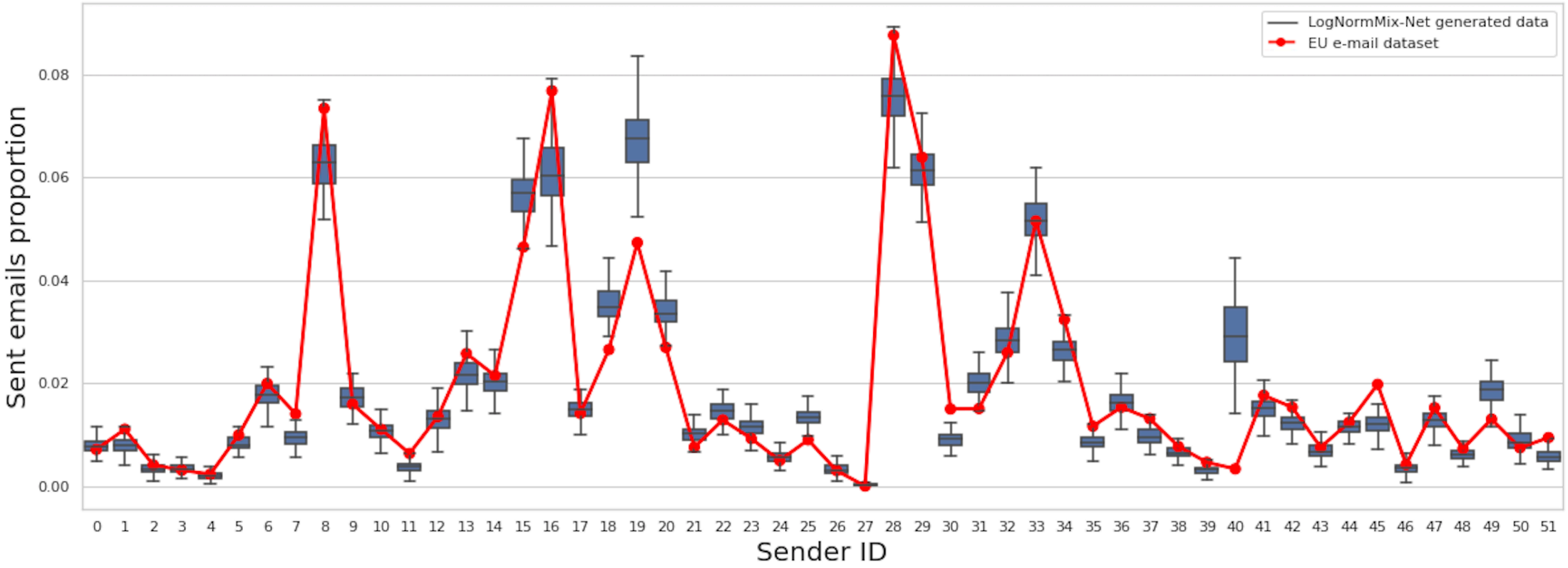








Sender outdegree



2. Generating multi-party e-mail threads

Leveraging pre-trained generative language models

Model Name	Number of Parameters
GPT-2	1.5B
GPT-3 Ada	2.7B
GPT-Neo	2.7B
GPT-3 Babbage	6.7B
GPT-3 Curie	13B
GPT-3 Davinci	175B

Table 1: Size of various GPT-based models. Source: Venturebeat:gpt-3s-free-alternative-gpt-neo

Goal: Simulate office network

- each individual has consistent, appropriate topics
- threads coherent and stay on topic

e-mail content generation

1. **e-mail subject line generation model**
   * fine-tune GPT-2 on the email subject lines from Enron corpus.
   * extract set of topic/keywords for each user (from their email subjects)
   * prompt the subject generation model with a sampled topic/keyword

Subject Personalization

Top 10 Keywords for user IDs 4, 5, and 10:

ID 4: ['Update', 'Meeting', 'Bullets', 'Weekly', 'Capacity', 'Pipeline', 'Storage', 'Project', 'Revised', 'List']
ID 5: ['Citizens', 'Agreement', 'Sale', 'Purchase', 'Deal', 'Contract', 'Option', 'City', 'Supply', 'Price']
ID 15: ['Meeting', 'Conference', 'Risk', 'Resume', 'Visit', 'Energy', 'Research', 'Power', 'Model', 'Summer']

ID 4

Pipeline News: August 26, 2001.
Pipeline Summary for October 11, 2001.
Meeting to discuss Team Selection -Reply.
Update on California Electricity Market.
Bullets 09/02/01.
Capacity Matrix Update.
Capacity Report.
Weekly Update from the Office of the Chairman.
Weekly Updates: Energy, Environment, and Weather.
Weekly Update on Power Markets & Energy Market.
Weekly Update - RTO Week -- Summary of Comments.
List of Accomplishments.

ID 15

Risk Management Simulation-Please review..
Summer and Fall Schedule, November.
Summer Intern Information.
Summer Associate Candidate - Angela Davis.
Summer Clerkship Program Winners.
Resume for Jeff Skilling.
Resume : Your Input Required.
Resume Submitted.
Power Point Presentation on Credit Risk.
Energy Analysis - New Issue.
Risk Systems Update for December 11th.
Visit to Portland - July 18.
Visit to Weather Desk.
Model Review Meeting - June 9, 2001.
Meeting in Portland - July 25.

ID 5

Sale of Napoleonville land.
Price and Interest Rates, as seen on the MarketWatch.
City of Mesa Update.
Contract or Training.
Sale of the Hines Hines/Nerdwood.
Agreement with PG&E.
Deal Request - M5B17.1.
Contract for: the Office of the Chairman, and for the oom.
Agreement with Drexel Energy.
Contract Payment Status Report.
Deal Correction Notice - Week of Oct 25.
Agreement with EPMI-ECI.

e-mail content generation

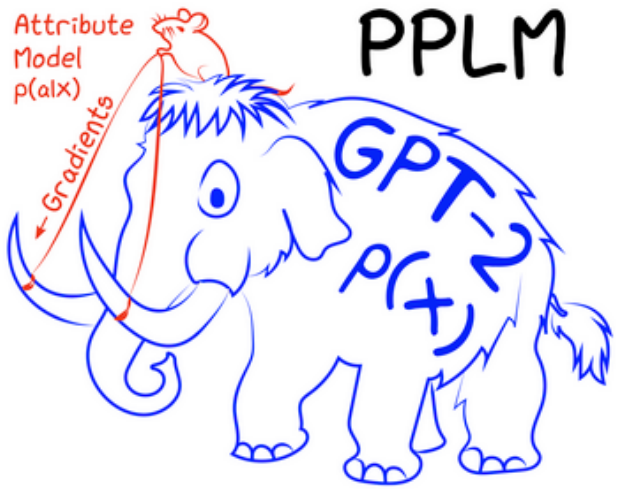
1. **e-mail subject line generation model**
   * fine-tune GPT-2 on the email subject lines from Enron corpus.
   * extract set of topic/keywords for each user (from their email subjects)
   * prompt the subject generation model with a sampled topic/keyword
2. **e-mail body generation model**
   * fine-tune on the email bodies from Enron corpus.
   * seed/prompt the generation with the email subject and existing email thread (if any)

Subject: Accountability Notice.	
New thread	**From:** Stephen Keller **To:** Amanda Lee; Mark Ford**;** Lauren Gould Hello all, Your Approval for this User is Important If you do not approve this User, you will receive a message from the user and will not be able to review your User ID and/or Account status. Please check the User ID and/or Account status in their profile to make sure their Approval is complete. Best, Stephen.
Reply e-mail	**From:** Mark Ford **To:** Stephen Keller **CC:** Amanda Lee; Lauren Gould Hello Stephen, It seems that your Approval is critical to have your team send you information on User ID and/or Account status so we can review and approve your User ID and/or Account status for you. We will need to know: how many days this User has been online, what percentage of his/her profile has been filled In the event this User has been online for more than one day, will he or she be eligible to receive all User ID and/or Account status? Regards, Mark.
Fwd e-mail	**From:** Mark Ford **To:** William Ware **CC:** Stephen Keller; Amanda Lee; Lauren Gould Hi William, Can you help out here?

Increasing the enticement of e-mails

Plug and Play Language Model (PPLM) [Dathathri et al. 2019]

- PPLM gives better control over the topic of the generated e-mails



- How can PPLM increase enticement?
  - cyber deception scenario: intruder searches file system for documents containing keywords of interest
  - add enticing keywords to the PPLM model's bag of words → encourages the model to use them during generation

Prompt: Meeting today at 10:30 am PDT. Hi Linda,	
Subject: Meeting today at 10:30 am PDT. **From:** Sandra Lynch **To:** Linda Perez	
Step size = 0.3	Hi Linda, We are meeting at the 10th-16th of November in San Francisco, CA to discuss the development on the East Coast of ENE. As such, the first meeting will be with Linda Adrienne Young and myself to discuss all the features of our plan, what we expect to see, the challenges, and what we can be grateful for in the next 2 weeks. Linda and his team will be meeting with the ENE developers and we will then move on to the 3-4 Meetings after our last meeting. I have attached a list of what we expect to see, and some of the features and opportunities in the 3-4 Meeting. Sandra.
Step size = 0.5	Hi Linda, You have a number of outstanding projects that we are working on. I would encourage all of you to contact us today to take a look at this project if you have questions. Meeting Attendees [IMAGE] [IMAG= E] [IMAGE] [IMAGE] [IMAGE][IMAGE] [IMAGE][IMAGE] [IMAGE] [IMAGE] Good Meeting Meeting Regards, Sandra.

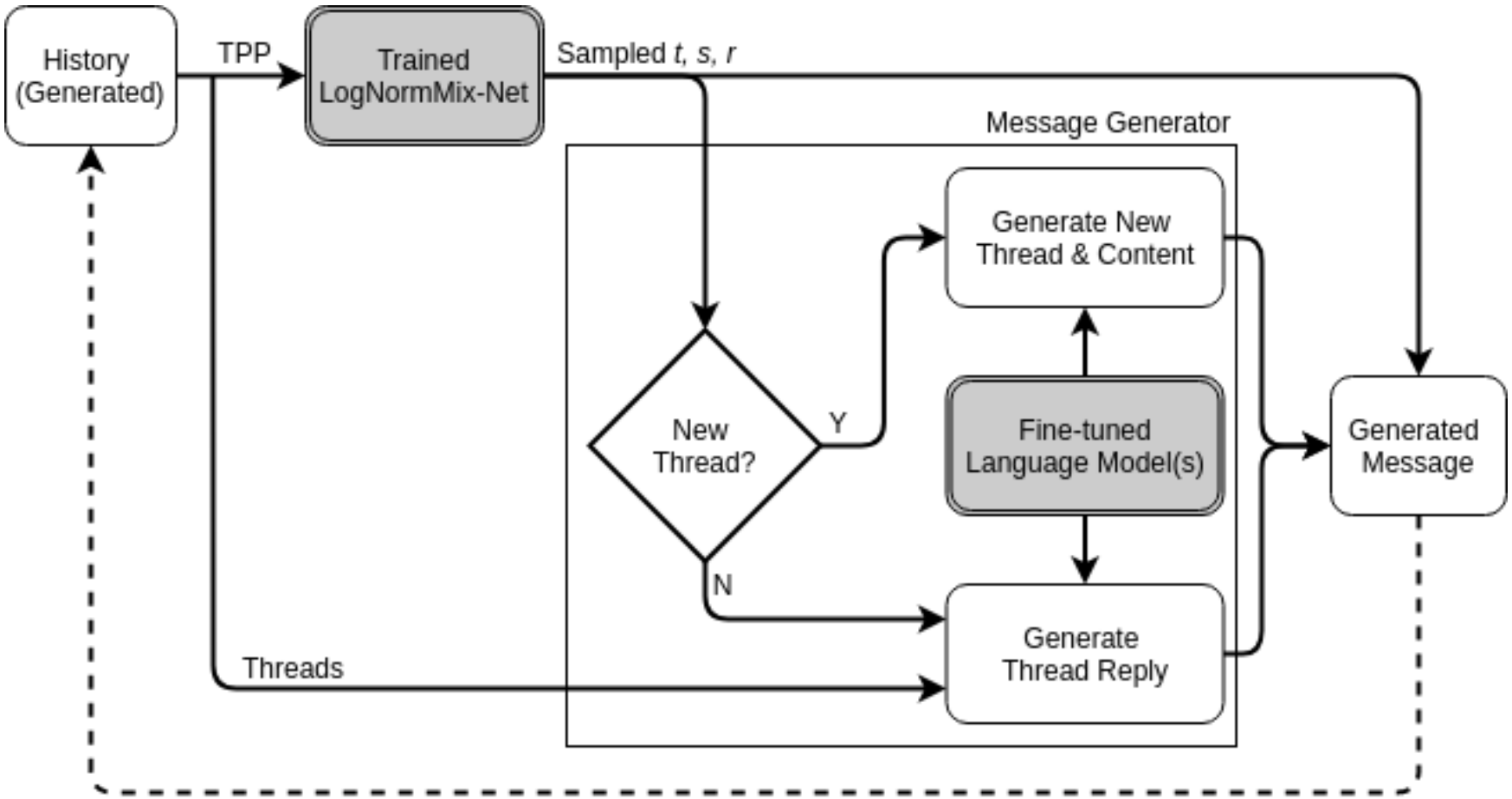
Evaluating the realism of generated content

Inspired by Karuna et al. 2018: *Enhancing cohesion and coherence of fake text to improve believability for deceiving cyber attackers*

- **Coherence:** similarity score between 2 consecutive emails within a thread (Google Universal Sentence Encoder + cosine-similarity)

- **Cohesion:** number of overlapping lemma types that occur in an email and its reply

Full Architecture



Challenges & Future Directions

- Privacy preserving generation of deceptive content
- Evaluating the realism and enticement of ML generated decoys
- Understanding/influencing the cyber adversary
- Quantifying the effectiveness of cyber deception
- Protecting our autonomous systems from being deceived

Thank You