# Leveraging Knowledge Graphs and Trans E Embeddings for Event Extraction from Email Data

This presentation explores how knowledge graphs and Trans-E embeddings can extract structured event information from email data, transforming unstructured communication into actionable insights.

By Gaurav Surtani & Somesh Bagadiya

# Why Email Data?

> "Why should it be so hard to find something on email? We have chatbots for everything but this?"

### Vast Unstructured Information

Emails contain rich details about events, tasks, and social interactions.

### Hard to Analyze Systematically

Traditional methods rely on brittle keyword searching or rule-based systems.

### Knowledge Graph Solution

Represents entities and relationships in structured format for analysis.

### Tracking Event

Currently Gmail only provides Promotions & Social Labels.

# System Architecture

### Data Ingestion

Process MBOX format email data from Google Takeout

### Parsing & Serialization

Extract metadata and content, from MBOX to Graph format

### Graph Database Loading

Populate Neo4j with nodes (Person, Email) and relationships

### Embedding Generation

Convert tuples of 'From' , 'To' and 'Event' nodes to embeddings

### Agent Interaction

Query, analyze, and interact with the email knowledge graph

# What Are We Extracting?

### Meetings

Invitations, scheduling discussions, acceptances/declines, follow-ups, and participants.

### Tasks & Deadlines

Assignments, status updates, reminders, responsible individuals, and projects.

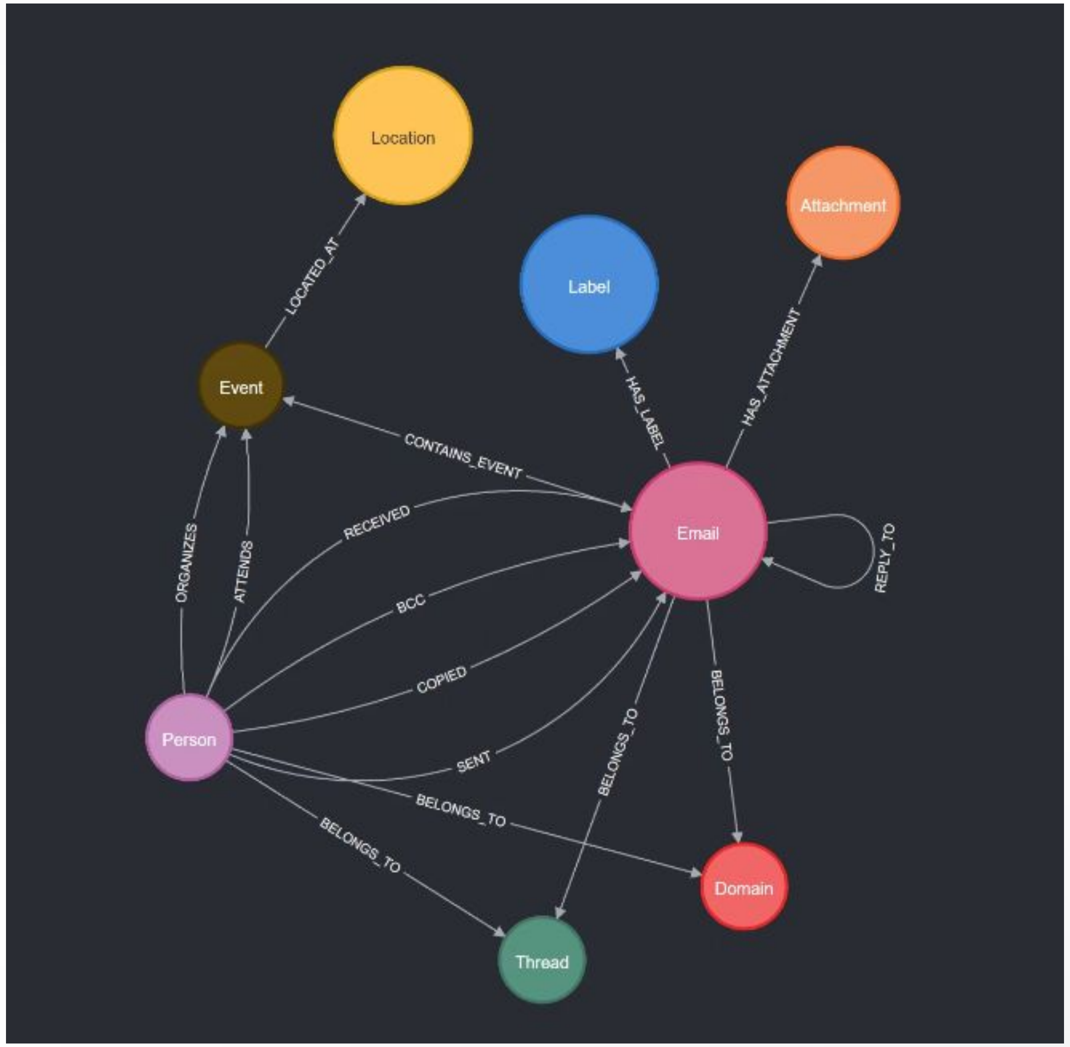
### Projects

Related email threads, key discussions, collaborators, and shared resources.

### Travel & Social Events

Itineraries, confirmations, invitations, and coordination efforts.

# Database Schema




| Focus Area | Why it's on the diagram |
|---|---|
| **1. Email = Hub** | Central node connects to *every* other entity. |
| **2. Person ← → Email edges** (`SENT`, `RECEIVED`, `BCC`, `COPIED`) | Encode *how* each person touched the message. |
| **3. Thread & Reply loop** (`BELONGS_TO`, `REPLY_TO`) | Reconstructs conversation trees. |
| **4. Context nodes** (Label, Attachment, Domain) | Add metadata richness. |
| **5. Event → Location path** | Extracts meetings and where they happen. |
| **6. Graph extensibility** | New nodes/edges can be added pain-free. |

# Knowledge Graph Components

neo4j

neo4j (default)

| | |
|---|---|
| **Nodes** | 21272 |
| **Relationships** | 59085 |
| **Labels** | 6 |
| **Relationship Types** | 8 |
| **Property Keys** | 11 |

# TransE: Translating Embeddings from Knowledge Graphs

## Core Concept

TransE models relationships as translation operations in the embedding space.

For a triple (head entity h, relationship r, tail entity t):

**$h + r \approx t$**

## Learning Process

TransE uses a margin-based ranking loss function.

It minimizes the "energy" for valid triples.

And maximizes the energy for corrupted (negative) triples.

# How exactly does Trans E work for  us?

2

3

4

## Setup

- Create a KG Triples dataset
- Head, Relation, Tail of the complete graph
- Setup model parameters,

## Training

- Set Epochs
- Each epoch over batches of (h, r, t)
- Calculate score over original data
- Then generates negative / corrupted triples by replacing entities.
- Computes the loss using Trans-E loss method

## Loss Minimization

- After each epoch,  loss function aims to increase the score of positive samples more than the negative scores.
- Then performs Backpropagation to recalculate gradients

## Normalization

- On each loss function update, it also updates the embeddings of the model.
- Essentially, it normalizes the entity embeddings so that vector lengths don't overgrow, which is a known issue of TransE.
- Logging Average loss for each epoch.

**Algorithm 1** Learning TransE

**input** Training set $S = \{(h, \ell, t)\}$, entities and rel. sets $E$ and $L$, margin $\gamma$, embeddings dim. $k$.

**initialize** $\boldsymbol{\ell} \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each $\ell \in L$
  $\boldsymbol{\ell} \leftarrow \boldsymbol{\ell} / \|\boldsymbol{\ell}\|$ for each $\ell \in L$
  $\mathbf{e} \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each entity $e \in E$

**loop**
  $\mathbf{e} \leftarrow \mathbf{e} / \|\mathbf{e}\|$ for each entity $e \in E$
  $S_{batch} \leftarrow \text{sample}(S, b)$ // sample a minibatch of size $b$
  $T_{batch} \leftarrow \emptyset$ // initialize the set of pairs of triplets
  **for** $(h, \ell, t) \in S_{batch}$ **do**
    $(h', \ell, t') \leftarrow \text{sample}(S'_{(h,\ell,t)})$ // sample a corrupted triplet
    $T_{batch} \leftarrow T_{batch} \cup \{((h, \ell, t), (h', \ell, t'))\}$
  **end for**
  Update embeddings w.r.t. $$\sum_{((h,\ell,t),(h',\ell,t')) \in T_{batch}} \nabla \left[\gamma + d(\boldsymbol{h} + \boldsymbol{\ell}, \boldsymbol{t}) - d(\boldsymbol{h'} + \boldsymbol{\ell}, \boldsymbol{t'})\right]_+$$

**end loop**

**Source :** Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating Embeddings for Modeling Multi-relational Data. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 26). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf

# Use Cases of adding Trans-E to Email KG

## Event Identification

Classify email types and discover patterns - Is there an interview?

## Link Prediction

Event information - Who is the interviewer? Where to meet?

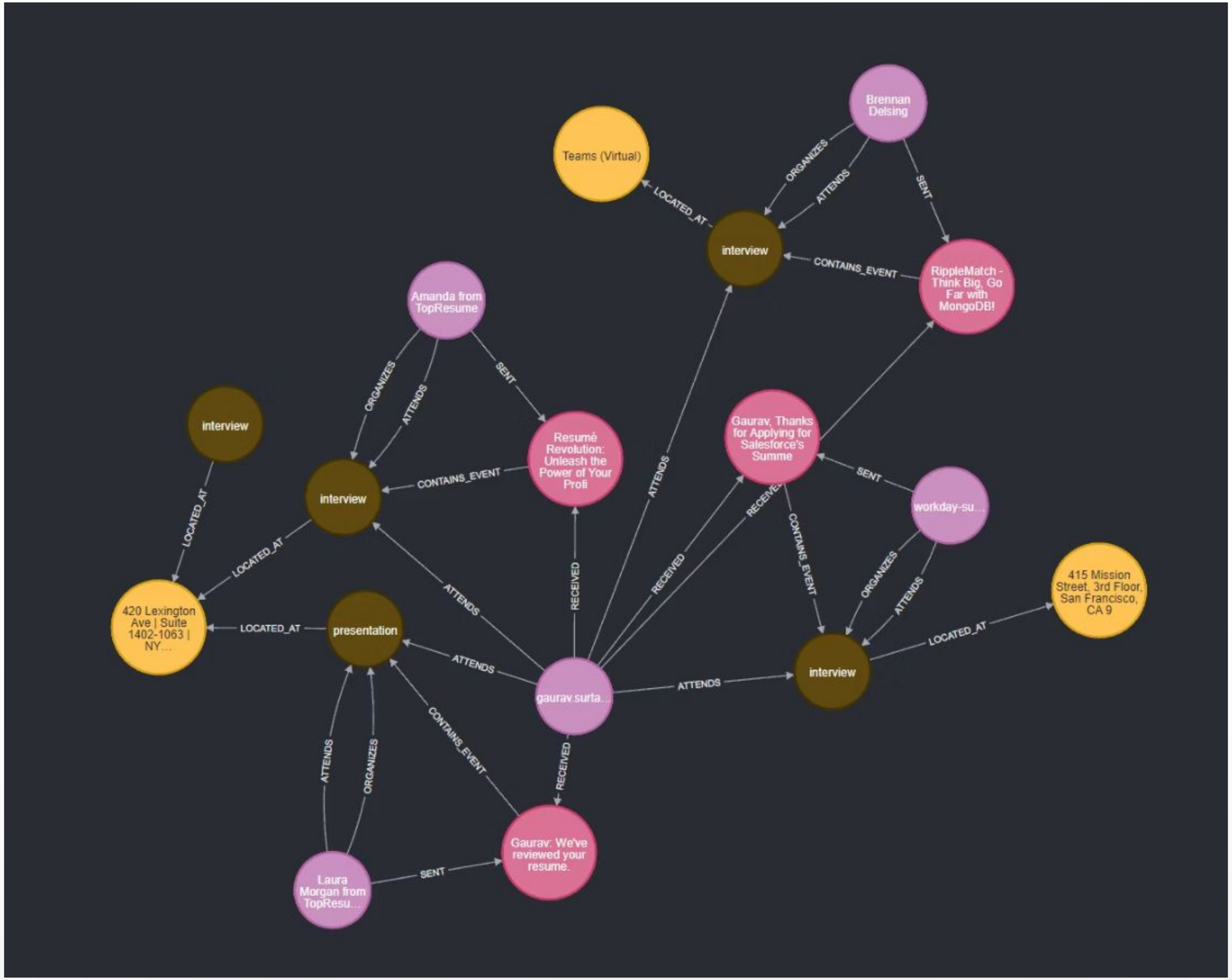
## Participant Discovery

Find missing event participants - Who are the other people involved in the interviewer, recruiter? Identifying helpful people?

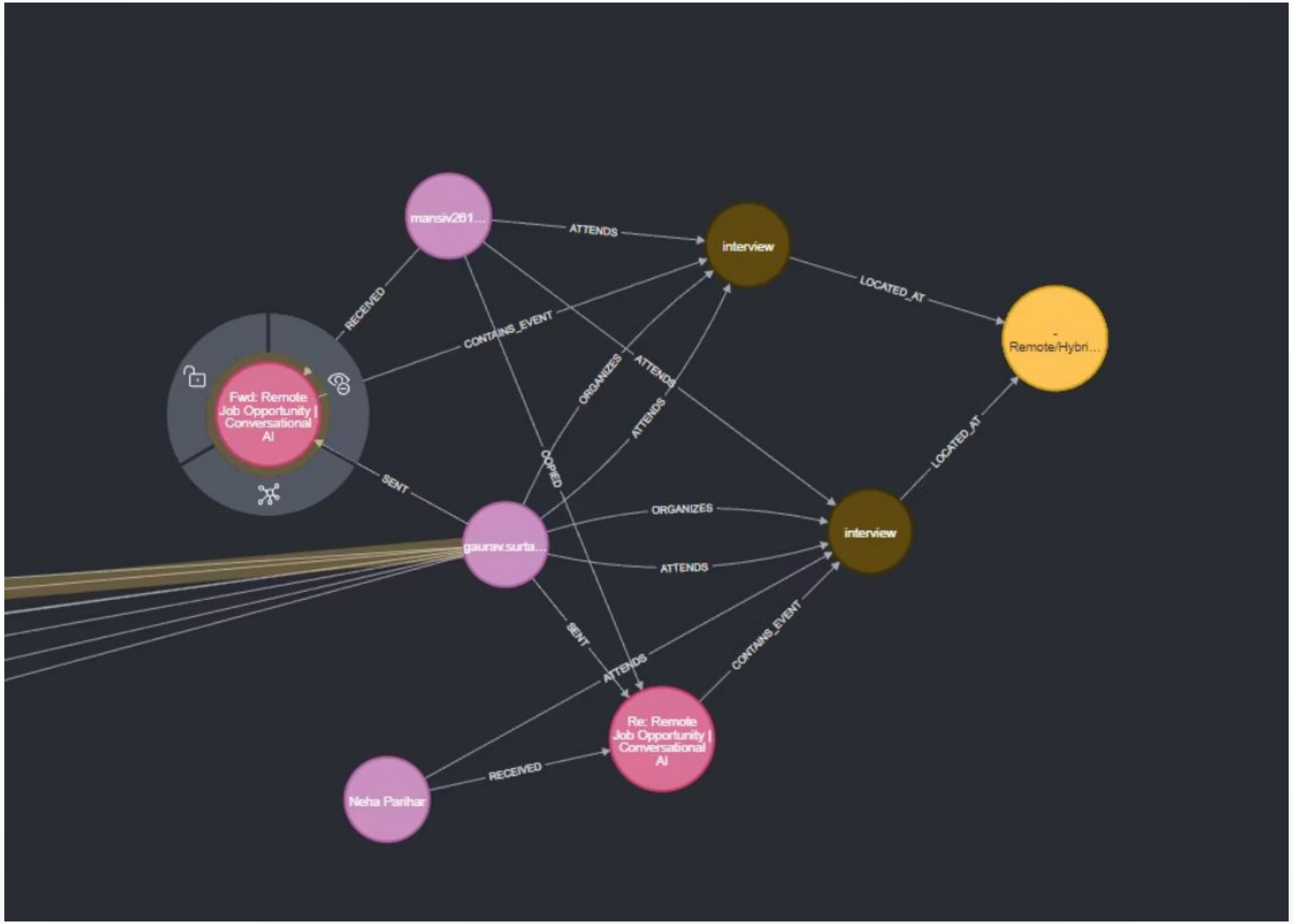
## Email Clustering

Group related communications - Other than Social , Promotions, we can have Flight Update, Meetings Tags.

# Examples (Events):

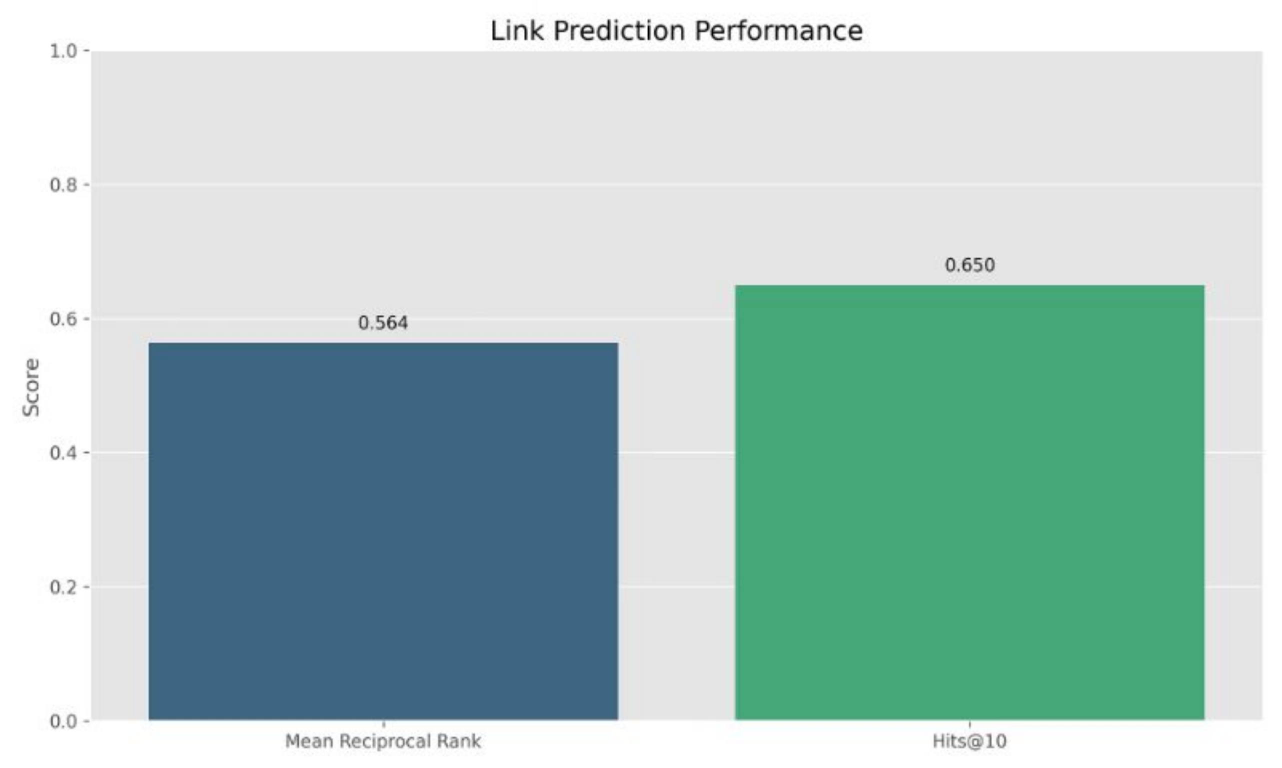


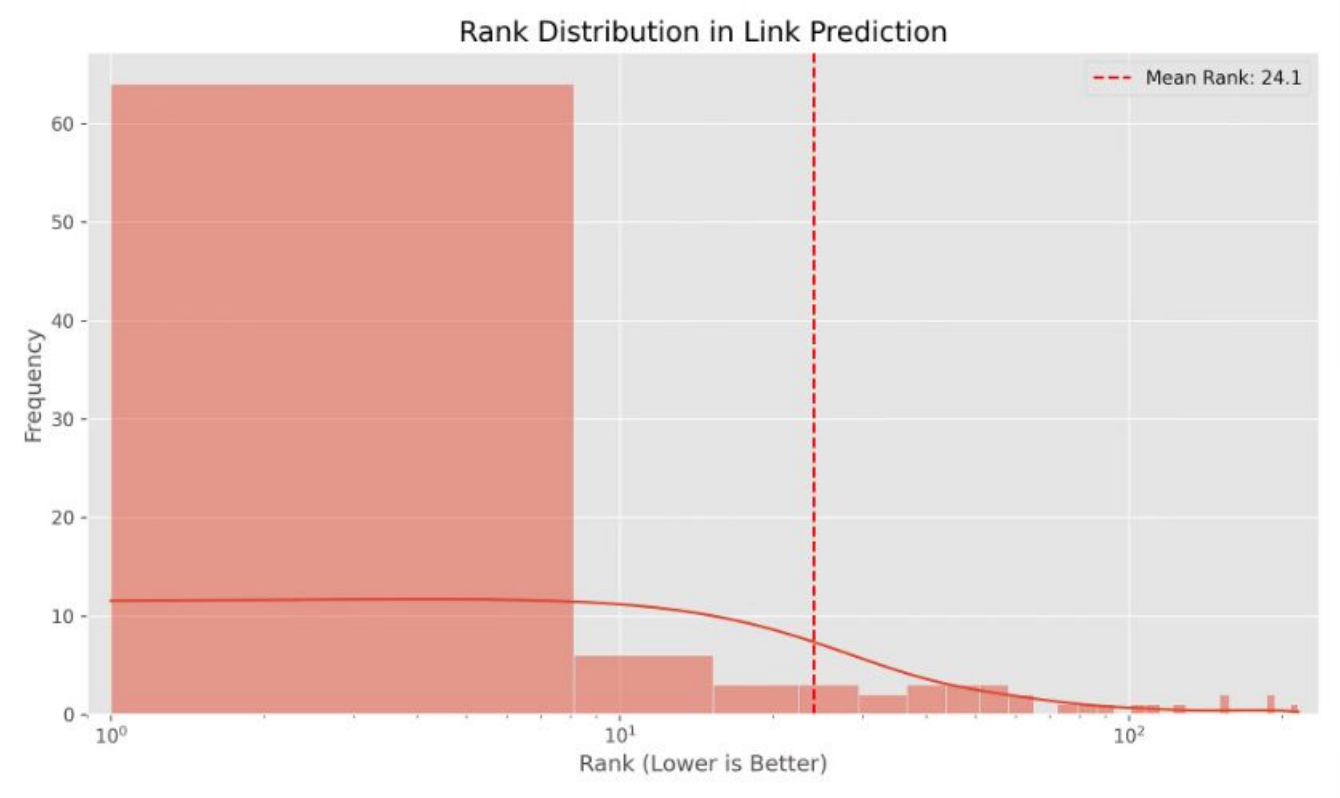
# Examples (Interview Event):



# Evaluation of TransE

## Link Prediction Performance



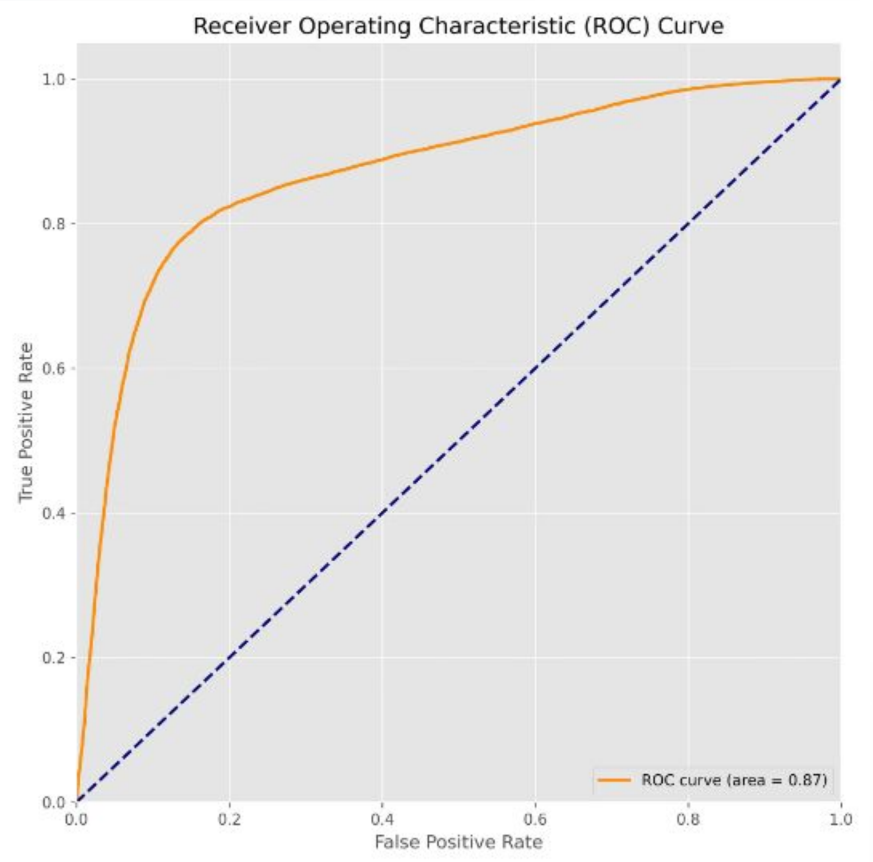



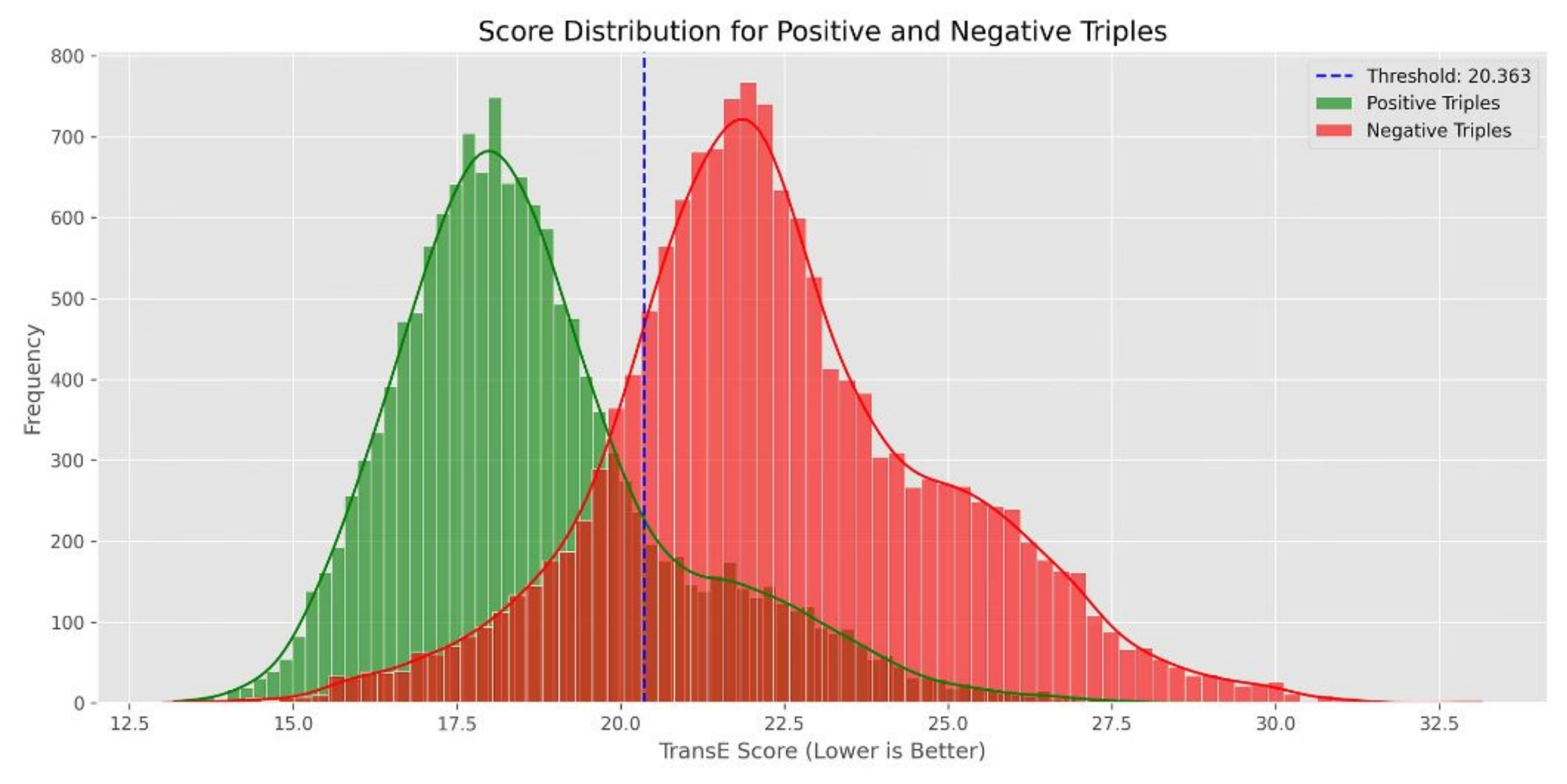

With an MRR of **0.564**, our model ranks the correct email recipient fairly high on average.

A **Hits@10 score of 65%** confirms that in most cases, the true recipient appears among the top 10 predictions — a strong result given the noise in email networks.

Most true links appear in the **top 10–20 ranks**, with a long tail of harder predictions.

The average rank of **24.1** shows that our predictions are not only accurate but also confidently placed near the top.




For AUC of **0.87**, the TransE-based model performs **significantly better than random** — this indicates that our learned embeddings are capturing meaningful structural information about email communication.

This separation validates that TransE embeddings **capture latent semantic relationships** in email communication, making it suitable for link prediction tasks.

# Advantages of Our Approach

### Structured Foundation

Explicit, queryable structure representing communication entities and relationships.

### Pattern Discovery

Uncovers non-obvious similarities within graph structure.

### Inferential Capabilities

Goes beyond directly observed connections to predict missing links.

### Data-Driven Approach

Adapts to the nuances of specific email datasets.

# Challenges & Limitations

## Graph Sparsity

Smaller datasets make learning robust representations difficult

## Event Representation

Bridging low-level relationships to high-level concepts

## Scalability

Large graphs require significant computational resources

## Model Limitations

TransE struggles with complex relationship patterns

# Future Work

## Richer Graph Schema

Include nodes and relationships more closely related to events

## Temporal Embeddings

Incorporate time information directly

## Advanced KGE Models

Explore TransH, RotatE, ComplEx, DistMult

## Hybrid Approaches

Combine embeddings with rule-based systems

# Conclusion

## Structured Approach

We've built a system that transforms email data into a knowledge graph.

## Vector Representations

TransE embeddings capture latent patterns in communication networks.

## Powerful Methodology

Our approach moves beyond traditional methods to uncover complex event patterns.