Machine Learning on GDPR-Compliant Data: Link Prediction in an Enterprise Social Network

Rita Korányi, José A. Mancera, and Michael Kaufmann. Lucerne University of Applied Sciences and Arts (HSLU)
Contact: rita.koranyi@hslu.ch

ABSTRACT

The amount of available information in the digital world is way larger that people are able to consume. Beekeeper AG provides a social network platform for frontline employees, who are typically not having permanent access to digital information. Finding the relevant information for being able to perform their job requires efficient filtering principles, to reduce the time spent with search and save work hours.

We conducted an experiment with GDPR-compliant user interaction data from the platform. GDPR restricts the access to texts, pictures or other kind of content of the user interaction. Thus, we modelled the data as graphs, used the graphs structural properties as features, and compared two predictive models with machine learning algorithms to extract and predict network patterns among users.
The research results contribute to a scientific understanding of online and offline information flow within an enterprise social network and provide empirical insights into how employees are currently connected and may be connected in the future.

METHODS

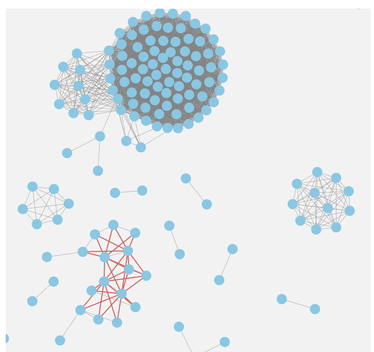
1. **Data Sampling**
   - Q1 and Q2 2019 for exploratory analysis and artifact 3 for model training and validation (appr. 15 mio. records each)
   - 280.000 records loaded into Neo4j (CW1 2019)
2. **Exploratory Analysis**
   - Exploring user behavior patterns
   - Statistical and network analysis
3. **Artifact Design**
   - **Artifact 1:** Knowledge Graph – who knows whom
   - **Artifact 2:** Link Prediction with Neo4j Graph Machine Learning Algorithms
   - **Artifact 3:** Link Prediction for User Relationships with scikit-learn

   The link prediction problem was defined as **binary classification**.
4. **Evaluation**
   - Comparing AUCPR, model training time
5. **Validation**
   - Labelling of user relationships by Beekeeper based on ground truth

RESULTS

Artifact 2

- In-database algorithms
- Features engineered for nodes
- Data treated as a graph
- AUPCR test data: **0.34**

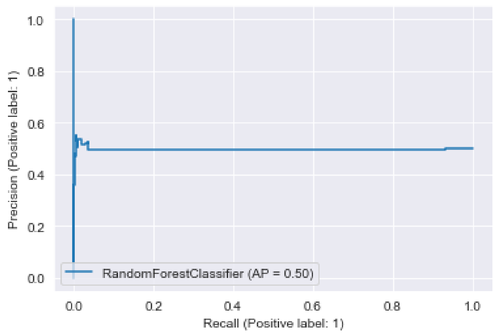


Artifact 1

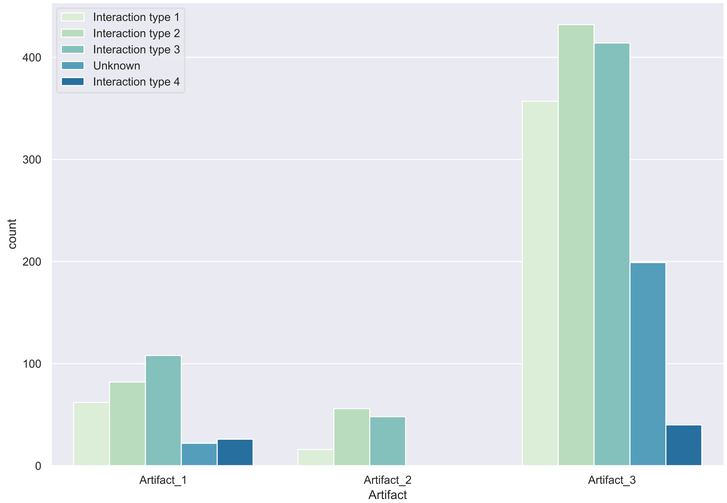
- Cypher query returns pairs of users
- Extracts **who knows whom**
- Definition of *know:* users had at least one chat interaction

Artifact 3

- Feature engineering with **NetworkX**
- Machine learning with **Scikit-learn**
- Features engineered for relationships
- Data is treated as tabular
- AUCPR test data: **0.50**



Validation Results of User Relationships per Artifact



Note: The interaction types have been generalized, the exact types are known to Beekeeper.

CONCLUSIONS

- Good approach to model **GDPR-compliant data as a graph** and mining graph features for machine learning
- Model performance not good, but the validation showed that they provided still **valuable business insights**
- Supporting **employee mental health** and avoiding overload at workplace
- Improving employee retention by **tackling social isolation** within the company
- The obtained network structure is **time dependent**, capturing accurate network structure is challenging

HSLU Hochschule Luzern

BEEKEEPER