OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

INF FAKULTÄT FÜR INFORMATIK

VAST Challenge 2020
Mini-Challenge 1: Graph Analysis

Supervisor: Prof. Dr.-Ing. Bernhard Preim
Dr.-Ing. Monique Meuschke
M.Sc. Uli Niemann

Presenter: Seyed Behnam Beladi
Atrayee Neog
Xiongjun Wang

Data Understanding

- There are 123,892,863 records.
- 5 Node type:
  1. Person (used in all channels, only nodes with a spatial location assigned)
  2. Product category (for the procurement channel, eType = 3)
  3. Document (from the co-authorship channel, eType = 4)
  4. Financial category (from financial demographics channel, eType = 5)
  5. Country (from the travel channel, eType = 6)

- 7 Edge type (eType): Edges always go from node type 1 to some other node type.
  0. Email
  1. Phone
  2. Sell (procurement)
  3. Buy (procurement)
  4. Author-of
  5. Financial (income or expenditure, depending on direction)
  6. Travels-to

Data Understanding

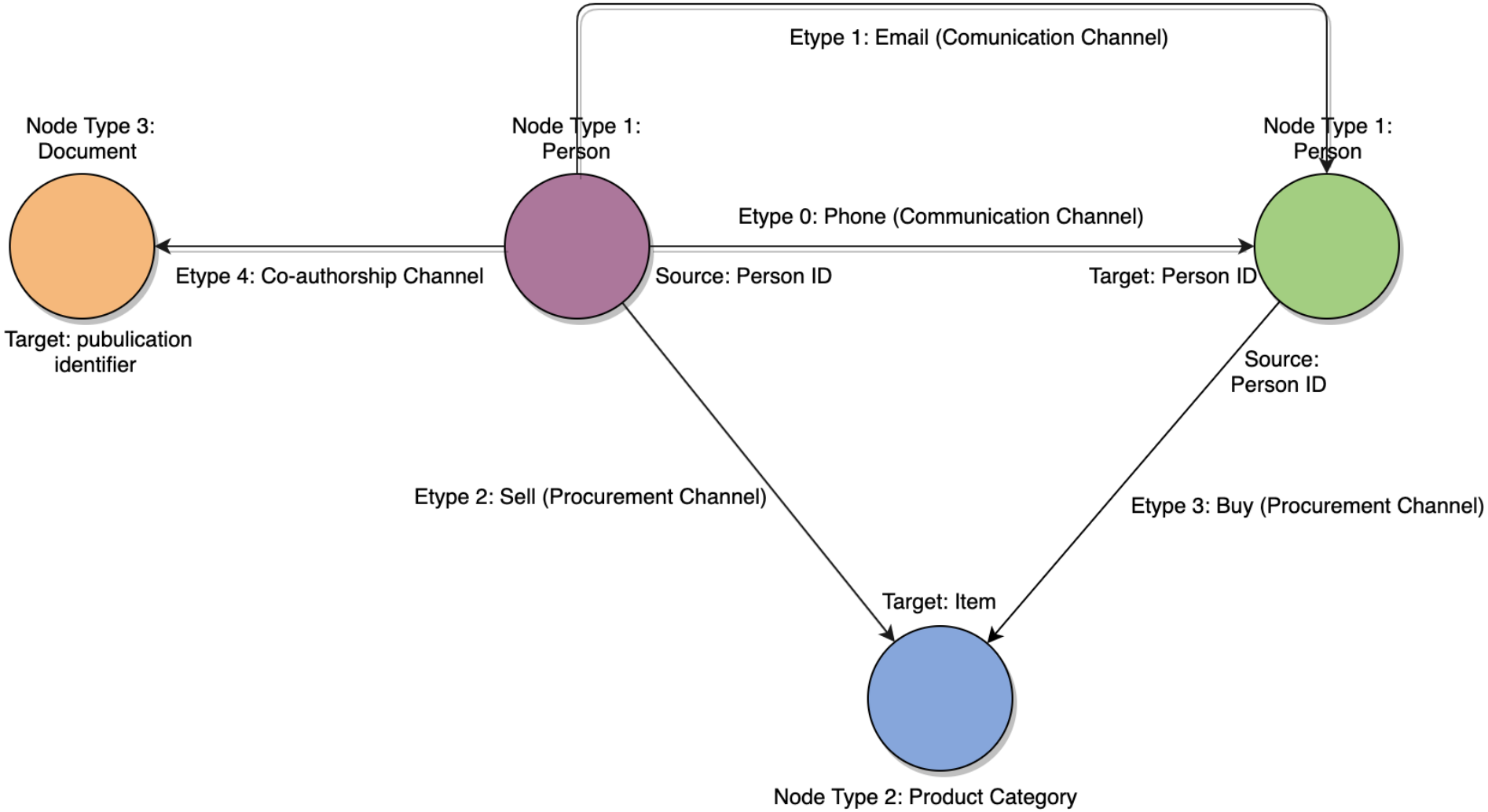
- There are 6 different channels of data, all of which are represented as a transaction between two nodes.

1. **Communications channels** (eType 0 and 1):represents direct connections between two persons.
   Source and Target columns are both person ID.
   Some records have location information, some don't.
   The weight for communications is always 1, representing 1 call or email.

- 2. **Procurement channels** (eType 2 and 3): Two people can be linked via the item they are both connected to.
   Source: Person ID, Target: item.
   The weight for procurements represents the value of the item.
   Procurements do not have location information.

3. **Co-authorship channel** (eType 4):represents publication of scientific or technical articles.
   Source column: Author (Person ID)
   Target column: publication with a unique identifier.
   The weight column indicates the fraction of the authors for the given publication.
   Authorship does not have location information.

Data Understanding

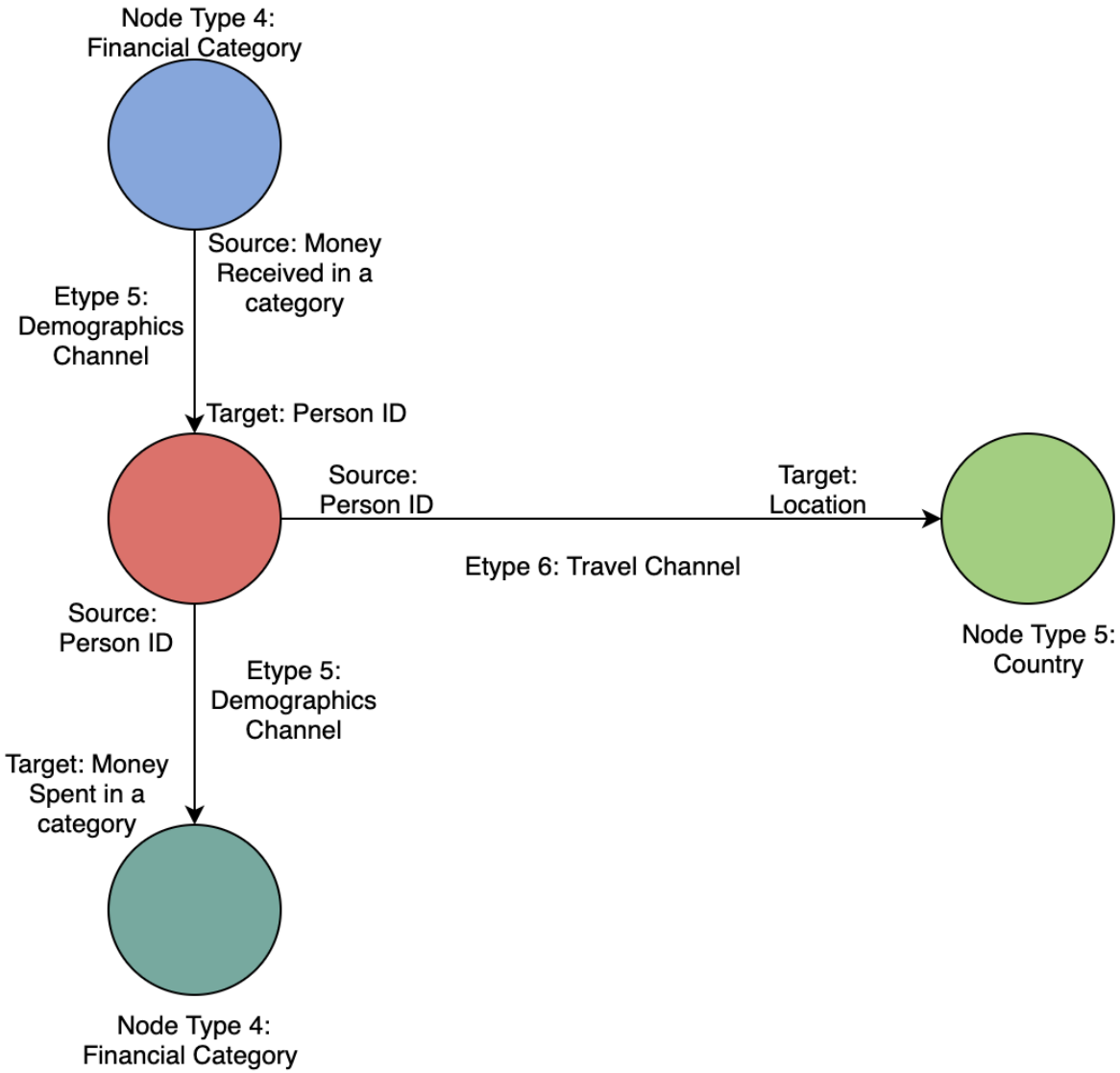


Data Understanding

4. **Demographics channel** (eType 5): represent the spending characteristics of each person in up to 30 categories, which are listed in the file DemographicCategories.csv
   Expenses: Source is person ID, Target column lists the money is spent in a category
   Income: Source column lists the money is received in a category, Target is person ID
   Time for all records in this channel is 31536000
   The weight channel shows how much is spent (or received) in a given category.
   Demographic records do not have location information.

5. **Travel channel** (eType 6): connects people (source column) with locations (target).
   Time: the start of a trip.
   Weight: length of the trip in days.
   All location columns should have data for each record.
   The SourceLocation and TargetLocation columns: countries of the origin and destination
   More specific latitude and longitude values are also provided.

Data Understanding



Data Understanding

data

Search

Name	Date Modified	Size	Kind
CGCS-GraphData-NodeTypes.csv	Mar 5, 2020 at 3:03 PM	2 MB	CSV D
CGCS-GraphData.csv	Apr 7, 2020 at 11:00 PM	6.35 GB	CSV D
CGCS-Template-NodeTypes.csv	Apr 2, 2020 at 1:49 PM	664 bytes	CSV D
CGCS-Template.csv	Apr 2, 2020 at 1:43 PM	39 KB	CSV D
DemographicCategories.csv	Apr 2, 2020 at 1:34 PM	735 bytes	CSV D
NodeTypeDescriptions.csv	Apr 2, 2020 at 2:46 PM	261 bytes	CSV D
Q1-Graph1.csv	Apr 1, 2020 at 9:57 AM	51 KB	CSV D
Q1-Graph2.csv	Apr 1, 2020 at 10:43 AM	55 KB	CSV D
Q1-Graph3.csv	Apr 1, 2020 at 5:16 PM	30 KB	CSV D
Q1-Graph4.csv	Apr 1, 2020 at 5:17 PM	31 KB	CSV D
Q1-Graph5.csv	Apr 1, 2020 at 9:24 PM	17 KB	CSV D
Q2-Seed1.csv	Apr 1, 2020 at 9:35 AM	168 bytes	CSV D
Q2-Seed2.csv	Mar 28, 2020 at 9:12 PM	169 bytes	CSV D
Q2-Seed3.csv	Apr 1, 2020 at 11:19 AM	160 bytes	CSV D

- The BIG graph: All records collected by CGCS are contained in a single file (CGCS-GraphData.csv). There are 123,892,863 records in this file. The uncompressed size is 6.2 GB.
- A template file (CGCS-Template.csv) is provided in the same edge list graph format as the large graph data. The template is a profile of activities that CGCS has built to represent suspicious activity associated with the hack. CGCS researchers hope that the group responsible will match, or partially match, this graph pattern.

Data Understanding

- Files have been provided for you to easily identify the node type of any unique identifier in the data. See CGCS-GraphData-NodeTypes.csv for node types in the large graph and subgraphs that have been extracted from it. See CGCS-Template-NodeTypes.csv for node types in the template.

- Candidate Subgraphs: Five subgraphs are provided for comparison to the template in the Question 1. They are: Q1-Graph1.csv
  Q1-Graph2.csv
  Q1-Graph3.csv
  Q1-Graph4.csv
  Q1-Graph5.csv

- Seed Graphs: Three seed graphs are supplied as starting points for your search in question 2. The seed files are: Q2-Seed1.csv
  Q2-Seed2.csv
  Q2-Seed3.csv

Vielen Dank für Ihre Aufmerksamkeit!

www.ovgu.de