Detecting social relationship clusters using graphs

Alberto Iriarte
Analytical Tools and Big Data Projects | CaixaBank

alberto.iriarte@caixabank.com
INDEX

► Introduction
► Why choose a Graph Database?
► Spatial & Graph Project at CaixaBank
  ► Architecture
  ► Data
  ► Integration & data exploration
► Use Case: Analyzing social relationships in customer data using graphs
► Other Use Cases in CaixaBank
INTRODUCTION: BIGDATA PROJECT

CaixaBank, one of the main banks in Spain, started in 2014 its great digital transformation project to become a “data-driven” entity.

The main milestones of CaixaBank’s digital transformation project were the following:

- **2014**: “Kick Off” digital transformation project (aka BigData Project) with Oracle as a main technology partner.
- **2015/16**: Evolution of the main data repository (DataPool), Security, Ingest, Governance and Data Quality. First industrializations with R, ODI (ETL).
- **2017**: Consolidation of the ecosystem of Analytical Tools (OBIEE, QlikSense, R, ...) for DataScience and business users (self-service).
- **2018**: Augmented Analytics: AI, ChatBots. NRT. New tools (DataRobot, Jupyter...).
- **2019**: Launch Spatial & Graph project (PGX). ML/Models execution platform for DS. Explosion of Self-Service.

CUSTOMERS 15.7 million
WHY CHOOSE A GRAPH DATABASE?

While relational databases are very useful for storing tabular data that fits into a pre-defined schema of rows and columns, they are not very efficient to find interconnections within a data set.

Forcing a highly connected data set into a relational database commonly results in severe performance issues in query return time.

As a result, a new type of database, called a graph database, has emerged to store relationship-oriented data naturally and efficiently using nodes and edges.

KEY POINTS

► Flexible schema: we can have nodes and edges with different number of attributes.
► Data and relationships are equally important
► It allows to run native graph algorithms (out of the box or custom).
► PGQL simpler than PL / SQL when it comes to looking for relationships at various levels of depth
► Improve the scalability and performance problems of relational databases. Linear computational cost.
► Oriented to analyze the relationships between objects: More suitable for modeling recommendation systems, route calculation, fraud detection, social relationships
Our graph infrastructure consists of three different parts:

1. Preparing edge/node data for PGX in EXADATA.
2. Loading the data into the Graph Server and generating the graph in memory.
3. Visualization of the graph and Batch/Online integration.
The **Graph Server** runs on a BDA appliance composed of six nodes with the following distribution:

**PRE**
- dpsgr01node01.lacaixa.es
  - 256 GB
- dpsgr01node02.lacaixa.es
  - 256 GB
- dpsgr01node03.lacaixa.es
  - 512 GiB

**PRO**
- dpsgp01node01.lacaixa.es
  - 256 GB
- dpsgp01node02.lacaixa.es
  - 256 GB
- dpsgp01node03.lacaixa.es
  - 1.5 TB
- dpsgp02node02.lacaixa.es
  - 512 GB
- dpsgp02node03.lacaixa.es
  - 256 GB
- dpsgp02node01.lacaixa.es
  - 256 GB
The volume of data loaded in the graph contains 400 millions of Nodes and 2.000 millions of Edges

Historical Data:
- 12 months of money transfers
- 6 months of credit card transactions

Nodes (400 millions)
- Edges (2.000 millions)
AVAILABLE INFORMATION

There are around 100 attributes available in the main graph.
GRAPH & DATA VISUALIZATION

We developed a graph database browser to allow technical and business users to explore the Graph. We also offer Batch integration for large-scale searches and online API for real time analysis of money transfers.

**Technical users** and **Data Science** with PGQL knowledge

**Massive analysis** to discover hidden relations between clients (by phone, address,...)

Discovery and self-service for **Business users** and analysts.

**Real time** analysis of money transfers
ANALYZING SOCIAL RELATIONSHIPS IN CUSTOMERS

The idea behind this project is to enrich and complete the lack of information (income, risk rating,..) of some of our clients based on their social relationships with other clients that we know better.

“Tell me who you go with and I’ll tell you who you are.”
SOURCES OF INFORMATION

Connections between clients are generated based on 3 different sources of information to detect links between clients from the point of view of “social relationships”.

Credit Card Payments at the “same time/place”
- Create connections between two clients based on a purchase in the same store with less than 3 minutes of difference.
- Use of 12 months of credit card purchases historical data in restaurant and leisure businesses.
- Delete connections that do not match a minimum of days and/or in a minimum of different shops.

Money Transfers with “social” purpose
- Create connections between two clients if they have made a money transfer between them.
- Use of 12 months of money transfers historical data.
- Delete connections that do not contain expressions referring to gifts, events, trips, family, etc in the purpose field.

“Bizum” Payments
- Create connections between Bizum users who send money to each other.
- Use of 12 months of Bizum historical data.
- Delete connections that do not match a minimum of transfers per year.
A preliminary analysis of the data sources reveals the following connections between clients:

**Total Clients:** 5,364,931 (30%)

- *linked with one or more source of data*

- **Credit Card Payments:** 569,789 clients
- **Money Transfers:** 4,989,546 clients
- **Bizum:** 1,446,518 clients

**Connections:**

- Only Credit Card Payment: 76,895 clients
- Credit Card and Money Transfer: 177,291 clients
- Credit Card and Bizum: 30,814 clients
- Money Transfer and Bizum: 863,239 clients
- Only Money Transfer: 3,664,227 clients
- Only Bizum: 267,676 clients
- All sources: 284,789 clients

**Distribution:**

- 74.7% - 1 source
- 20.0% - 2 sources
- 5.3% - 3 sources
The goal is to create a subgraph based on the already created main graph, loading only the nodes and edges necessary to identify the social communities.

- **Nodes (400 M)**
- **Edges (2.000 M)**
CREATE A NEW GRAPH FROM “ETL” + “RAW DATA”

Another approach is to create the Social Graph from “scratch” using an ETL process and raw data from the Data Repository. The resulting graph is smaller because we need only to create one node (client) and three edges.
ALGORITHM SELECTION: SLPA

Once we have defined the social graph (nodes and edges), the next step is select the best algorithm to identify the communities.

The two main requirements to select the algorithm are performance and overlapping community detection (modularity).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Modularity</th>
<th>Performance (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPM</td>
<td>0.503</td>
<td>0.99</td>
</tr>
<tr>
<td>DEMON</td>
<td>0.491</td>
<td>17.87</td>
</tr>
<tr>
<td>SLPA</td>
<td>0.588</td>
<td>25.57</td>
</tr>
<tr>
<td>wSLPA</td>
<td>0.577</td>
<td>28.63</td>
</tr>
</tbody>
</table>

Speak-Listen Propagation Algorithm (SLPA) is an extension of the native Label Propagation Algorithm (LPA) to discover overlapping community structures in a diagram.
SOCIAL INFLUENCE COMMUNITIES DETECTED

Once the process has finished, we observed that clients belong to more than one community. We can also identify who is the top “influencer” client in each community according to the centrality of this person inside the cluster.

<table>
<thead>
<tr>
<th>Total of clients in a cluster</th>
<th>2,9MM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total of communities</td>
<td>6,1MM</td>
</tr>
<tr>
<td>Clients per community</td>
<td>4,4</td>
</tr>
<tr>
<td>Communities per client</td>
<td>9,1</td>
</tr>
<tr>
<td>Average friends</td>
<td>13,7</td>
</tr>
</tbody>
</table>
CONCLUSIONS

We can correlate the socioeconomic attributes of clients in the same community. This gives us the chance to enrich data for those clients that we don’t know very well, improving our predictive models.

Community “Superhero”

CREDIT CARD PROFILE
- Credit Card: 10%
- Debit Card: 40%
- Pre-paid Card: 50%
- Risk: Low

CONSUMPTION HABITS
- Avg. spending in restaurants: 200€
- Shopping habits: Online
- Financed Purchases: No
- Mobile: 80% Iphone

CAIXABANK SEGMENT
- Imagin
- Family
- Imagin – Family

DEMOGRAPHIC
- Age: 35
- Gender: M
- Education: High
- Income: >75K
- Occupation: Yes

INTERESTS
- Travel: Y
- Music: Y
- Fashion: N

FINANCIAL INVESTMENT
- Fixed Rent: 20%
- Variable Rent: 80%
- Loans: -
- Mortgages: -
- Renting / Leasing:-
OTHER USE CASES IN CAIXABANK

Cases of use that we are currently working to expand graphs technology to other business areas in CaixaBank:

Fraud detection: identify “mule bank account” that hackers use a intermediary account for fraudulent money transfers.

Predicting risk propagation: between interacting firms
THANK YOU!

alberto.iriarte@caixabank.com
linkedin.com/in/albertoil