Exploring Graph Data Science Methods in Neo4j

Farm Credit Services of America

Design Studio Release 6

1 Introduction

In this document, we wish to explore possible methods to analyze graph-based data with a focus on interpretation. The nature of loan-provision services require some level of understanding of the underlying metadata and relevant domain knowledge to truly understand and make informed decisions. This would allow Farm Credit Services of America to better utilize `NEO4J`.

Hence, this proof of concept would describe the

1. Dataset used and Code
2. Description of Components and its Interpretation
3. Centrality Measures and what it means
4. Node Similarity Measures for Product Recommendation to be Explored

2 Dataset and Code

The dataset that we used for the following proof of concept is by querying the `NEO4J` database with the following code:

```
MATCH (c1:Customer)-[:HAS_FINANCIAL]->(f:Financial)<-[:HAS_FINANCIAL]-(c2:Customer)
MERGE (c1)-[:SHARES_FINANCIAL {id: toInteger(c1.P_CustKey)+toInteger(c2.P_CustKey)}]-(c2)
WITH
MATCH (c1:Customer)-[:SHARES_FINANCIAL]-(c2:Customer)
RETURN c1.P_CustKey as customer_one, c2.P_CustKey as customer_two
```

This connects customers that shares financials, which allows us to create a graph that relates different customers based on **shared financial ties**. This is helpful especially for understanding how risks are shared across different customers in a business unit (as we would define later).

The reason behind exploring risks and financials (as apposed to products) is because the measures introduced here are **relatively simple and interpretable measures of node importance**. There are other methods described in Section 6 where we discuss algorithms that can be used for product recommendation.

The relevant code as a Jupyter Notebook and the data file `financial_graph.csv` accompanied.

3 Components of Related Business Units

3.1 Weakly-Connected Components

To define our definition of a business unit, we first have to define a **weakly-connected component**.

Formal Definition - A weakly-connected component is a maximal subgraph of a directed graph such that for every pair of vertices , in the subgraph, there is an undirected path from to and a directed path from to.

Human Definition - If customer A can reach customer B somehow, A and B are in a weakly-connected component.

Figure 1: Weakly-Connected Components of Financial Graph

Figure 1 shows all the weakly-connected compoenents of the financial graph. There are a 156 number of components, with the largest component having 15 number of nodes. This means that the percentage of nodes that are within the largest connected component of this graph is 3.59%.

This is a **relatively disconnected graph** since most social-network graphs has $\approx 99\%$ of nodes within the largest connected component. The observation should make sense since most customers who share financials are usually connected via family ties or are just a part of a bigger business operation.

3.2 Business Units as a Unit of Observation

Based on the observations stated above, we define a **business unit** as all the customers in the same weakly-connected component. Hence, there are 156 business units in this graph.

Note that business unit in this context refer to a grouping of customers based on the rule stated. To see how each customer in a business unit relate one another, we require metadata and additional domain knowledge to make any inference since customers can be connected operations or family members or other relationships.

We also use the business unit as a unit of observation since we are assuming that those who do not share financials do not pose financial risks to each other.

4 Centrality Measures & their Interpretations

Centrality measures are algorithmic methods to provide a metric to **determine the position, role and importance of a node**. There are different centrality measures that provide different understanding since the **notion of importance is dependent on what we consider important**.

4.1 Degree Centrality

Formal Definition - The number of links incident upon a node over the theoretical maximum links.

Human Definition - How many links over all possible links.

This is the simplest measure of the relative importance of a node with respect to other nodes. **A node with a high degree centrality means that it is connected to a lot of other nodes**, which gives a good measure of how important the node is.

From a risk perspective, a node with high degree centrality directly exposes risks (and distribute risks which makes the whole less risky) to the connected nodes.

4.2 Eigenvector Centrality

Formal Definition - A measure of the transitive influence of nodes where relationships originating from high-scoring nodes contribute more to the score of a node than connections from low-scoring nodes.

Human Definition - Am I connected to other highly connected nodes?

The most famous form of an algorithm that uses this measure is Google's PageRank algorithm. Nodes who are connected to other nodes that are more influential would be assign a higher score.

From a risk perspective, **a high eigenvector centrality implies the strong influence of the node since it can spread and distribute risks to a wide part of the network**. This shows the importance of a node (like the main business line/operation) but it also implies that a contagion can occur. Contagion in this context means that if this line of business fails, the whole business unit could fail.

4.3 Betweenness Centrality

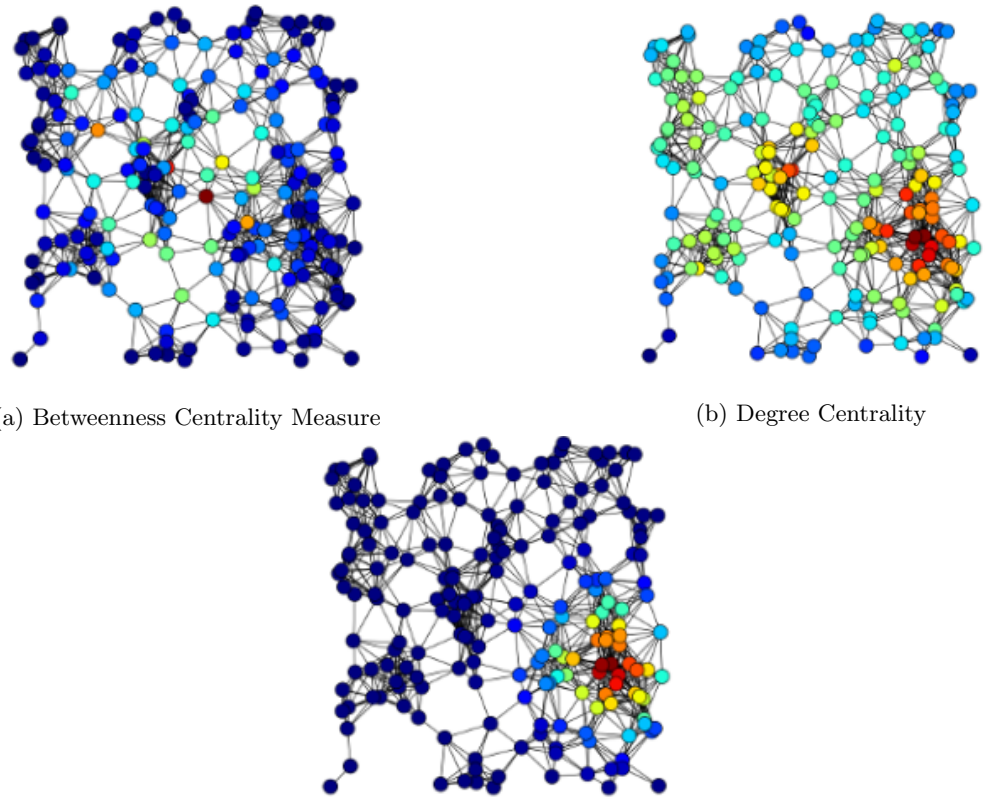
Formal Definition - Betweenness centrality measures the extent to which a vertex lies on paths between other vertices. Vertices with high betweenness may have considerable influence within a network by virtue of their control over information passing between others. They are also the ones whose removal from the network will most disrupt communications between other vertices because they lie on the largest number of paths taken by messages.

Human Definition - Am I a bottleneck (or star)?

Betweenness centrality measures a different kind of importance. While it does not imply its influence over the whole network, **a high betweenness centrality means that it exerts a strong influence on its direct neighbors**.

For an example, while the Farm Credit System is not the biggest financial service provider nationwide (low degree/eigenvector centrality) but can exert significant influence on the agriculture market (high betweenness centrality), assuming that the agriculture sector rarely borrows outside of the Farm Credit System.

4.4 Putting Things Together



(a) Betweenness Centrality Measure

(b) Degree Centrality

(c) Eigenvector Centrality

Figure 2: Different Centrality Measures on the Same Graph Reflecting Different Importance

Figure 2 shows the different centrality measures.

In Figure 2a, we see that the node with the highest betweenness centrality measure (the middle dark red node) is not considered important by other measures. However, it serves as a bottleneck (or star) that connects the two important components.

The eigenvector centrality in Figure 2c also differs from degree centrality in Figure 2b since it also accounts how connected the neighbors are. This is reflected in the cluster in the bottom-right corner where the eigenvector centrality has the largest measured value.

5 Case Study of a Business Unit

Looking at one of the business units obtained from the financial graph data. We observe some interesting patterns. Figure 3 provides a succinct visual interpretation of what the measure reflects and how we can understand this from a risk perspective as well as when considering providing a customer a loan.

5.1 Results from the Different Measures

Customer Key	Degree Centrality	Eigenvector Centrality	Betweenness Centrality
448256	0.1	0.047	0.0
5122	0.8	0.42	0.18
66023	0.5	0.309	0
66024	0.5	0.309	0
299688	0.4	0.243	0
111533	0.5	0.309	0
363503	0.3	0.170	0
61719	0.9	0.449	0.281
229913	0.6	0.275	0.215
924	0.5	0.309	0
229917	0.5	0.268	0.015

Table 1: Results from Applying the Measures (**Bolded Red Values** are the Ones In Future Discussion)

Table 1 shows the results from different centrality measure of this business unit. While Customer 61719 (in blue) has the highest measure for three centrality measures, we want to look at Customers 5122 and 229913 where they have a significant difference between degree and betweenness centrality. This would inform us on their different type of importance of the customer in the business unit.

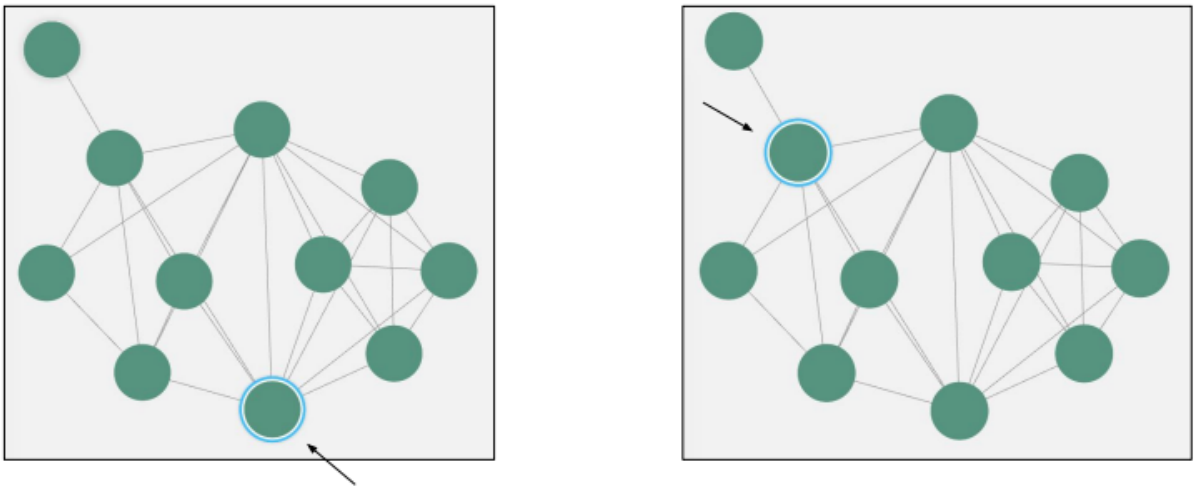
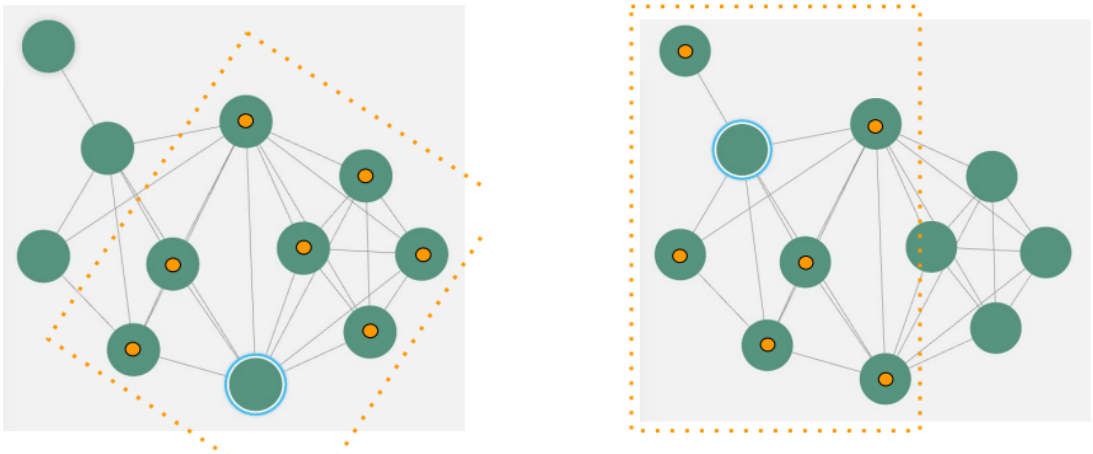


Figure 3: (Left) Customer 5122 has a higher eigenvector and degree centrality whereas (Right) Customer 229913 has a higher betweenness centrality

Note that from Figure 3, we can see that the highlighted node in the *right* frame is **connected to nodes that are less connected customers and one customer is entirely dependent on it**. On the other hand, on the *left*, **the node is connected to a lot of nodes that are strongly connected with each other**.

5.2 Interpreting and Inferring from the Measures



(a) Customer 5122 (b) Customer 229913

To make the example concrete, let us assume that this business unit raises cows and pigs, with cows being the central business. For Customer 5122 (who raises cows, or involved in raising cows like providing feed etc.) as shown in Figure 4a, it is **very well connected to other customers** who are involved in the cow raising business, and so **there is no dependence on a single customer within that group**.

On the other hand, Customer 229913 in Figure 4b might be the only supplier of feed for pigs, and the connected customers to 229913 have pigs as livestock. This means that they are **dependent on Customer 229913** though as a whole, they play a **smaller role in the business unit**.

5.2.1 Risks

With the connection of nodes being the relationship of *sharing financials with one another*, it is natural to first look at the problem from a risk perspective. In economic literature studying financial networks, sharing financial presents a risk (pun intended) and a reward:

1. **Risk** - Contagion, a very well connected and influential customer can cause other customers to fail if it fails, similar to a *domino effect*
2. **Reward** - Diversification, fluctuations in performance of a customer can be 'absorbed' by other customers

Referencing Customer 5122 in Figure 4a, with a lot of connections, **small changes** to one of the nodes **can not truly affect the whole network** since the risk is 'dissipated' across the network. In this scenario, *unsystematic risks* for Customer 5122 in Figure 4a is lowered as a result of the connection. However, any **severe shocks** on one of the nodes **can lead to a contagion**, where all other customers might face financial difficulties as a result of *systemic risks*.

On the other hand, the **likelihood of a contagion** for Customer 229913 in Figure 4b is **lower** since it is has relatively weak effect on the whole network. However, Customer 229913 **exposes a lot of unsystematic risks to its direct neighbors**, where it presents a bottleneck in the network.

5.2.2 Loan Provisions

For loan provisions, the idea behind this discussion stems from the **idea of contribution towards a financial statement** and we would be focusing on the discussion for situations like Customer 229913 in Figure 4b.

Customer 229913 has a 'star' shape within the network. Our intuition suggests that such patterns could arise if **Customer 229913 has a very good financial standing**, whereas the other customers that are

connected are simply **using its reputation to obtain more favorable loans**. This is analogous to Enron's financial statements and its 'special purpose entities' (*to be clear we are not accusing anyone of committing fraud*).

In the event that 'star' shaped financial relationships exists, it suggests that a **closer inspection in the individual contributions to the financial statement** has to be made to ensure that Farm Credit Services of America is not exposed to unexpected risks.

6 Node Similarity Measures

While we mostly described centrality measures and their interpretations, Machine Learning with graphs has other huge developments particularly in the unsupervised and semi-supervised methods. This is important because these developments would be able to help Farm Credit Services of America **recommend products to potential customers** based on certain characteristics that are important.

6.1 State of the Art Algorithms

A huge stride in Graph-based ML has been the ability to **measure the 'similarity' of nodes using unsupervised techniques**. The most famous of which are **DeepWalk, Node2Vec and Graph Neural Networks (and their spin-off architectures)**.

The three methods mentioned all **define the notion of node similarity based on shared neighbors** (nodes that are connected to one another, and on and on). They then **map this similarity to a vector space** where similarity measure such as the dot product or sin function can be applied

6.2 Collaborative Filtering

Node similarity is often used in the context of collaborative filtering, where we **recommend products to customers with some shared characteristics and connections**. In the context of Farm Credit Services of America, this means product recommendation and looking for potential leads.

6.3 Weaknesses of Such Methods

Although these methods have been successfully applied, they face the **weakness in interpretability**. This however, should not be too much of a problem for product recommendations since some recommendation is better than none at all. Furthermore, the mapping of vectors are based on similarity measures, which might not be applicable to measures of risks associations as discussed in the previous section.