Chocolate Bar 2020 Project

Yifei Dong

DSBA 6520 – Network Science

Jun 11th 2021

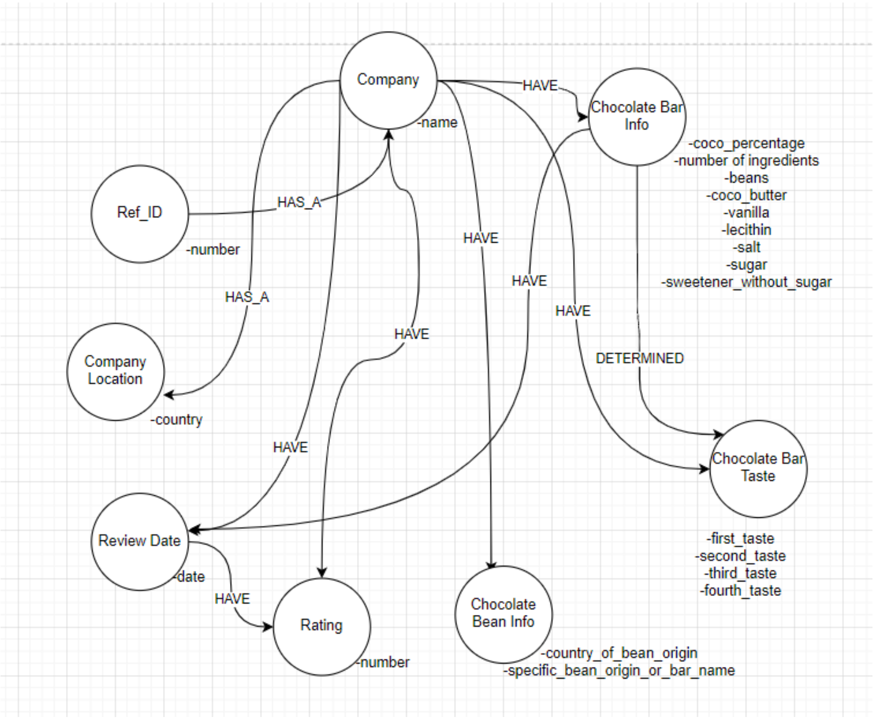
Professor Robinson

As a Chocolate shop manager at the South Park Mall, we import and sell chocolates from all over the world from hundreds of different companies. However, after Covid-19 has started, the sales have been declining. To provide a better shopping experience and great quality for our customers, I will use graph analytics to identify few questions below using the Chocolate Bar dataset:

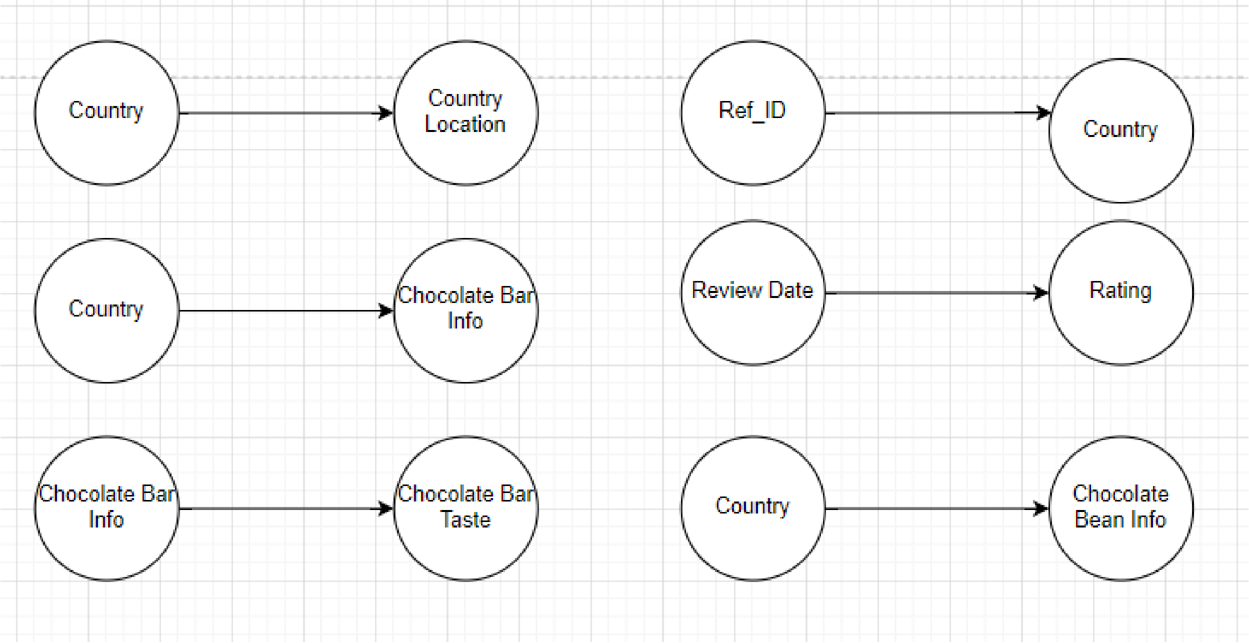
- Where are the best cocoa beans grown?
- Which countries produce the highest-rated chocolate bars?
- Which company has the highest rate?
- What is the most popular taste?

This dataset contains 21 columns and 2225 rows. The first question can help me find out where to purchase the best quality cocoa beans for my company to produce a high-quality chocolate bar. The second and third questions can help me identify which country and company I should import my product from that are most populated and received the highest rating from customers. The last question can help me determine what kind of variety my chocolate bar shop should have to target a wider range of customers. I will use the dataset to answer these questions and determine what chocolate will be purchased from which company in the next year.

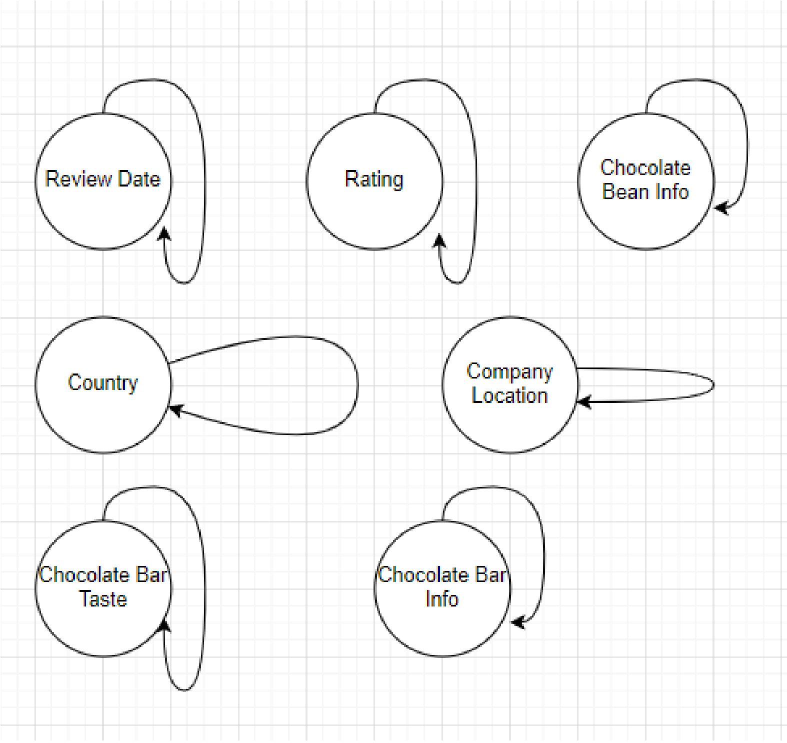
Graph data model



Bi-partite graph

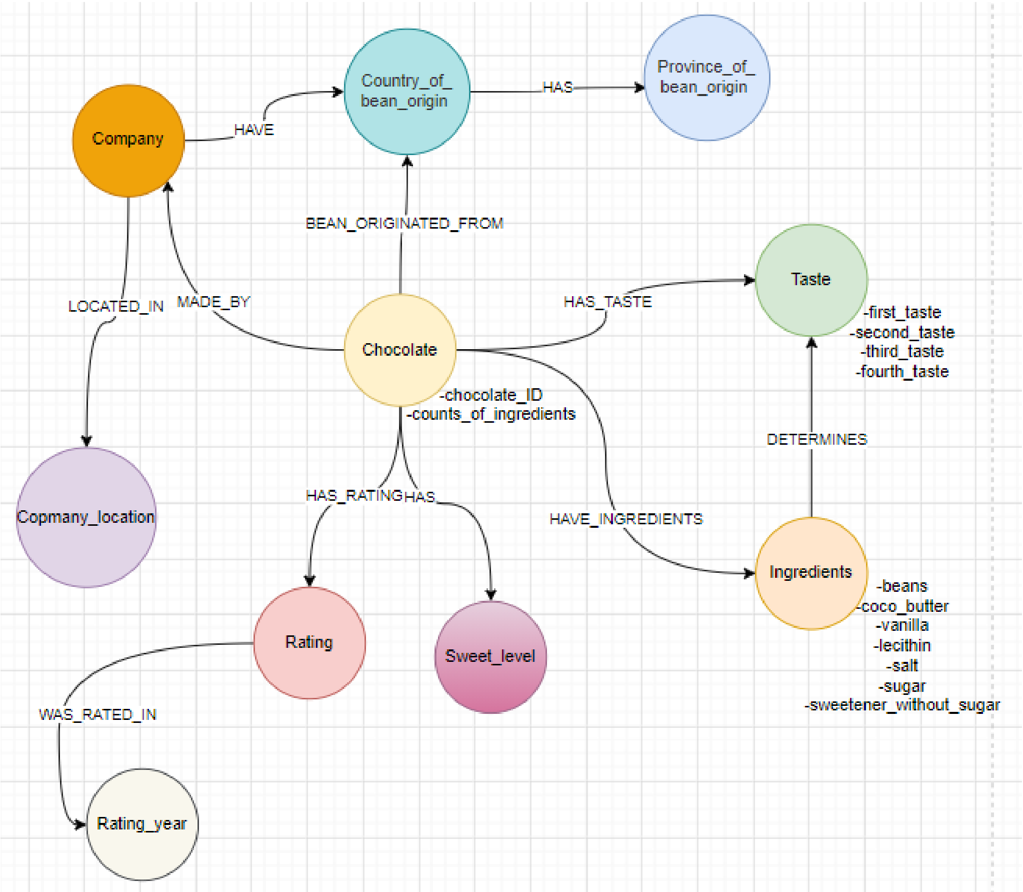


Mono-partite graph



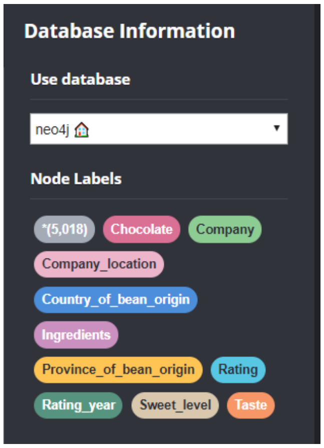
Part II

The updated version of the Data graph model:



The reason why I updated my graph model this way is because I grouped several ranges of cocoa percentage into sweet levels for better understanding. Ingredients and Taste nodes contain unique values that are connected with the chocolate node through chocolate ID. I have also connected company with country of bean origin for algorithm purpose.

Neo4j Database Setup and Screenshots



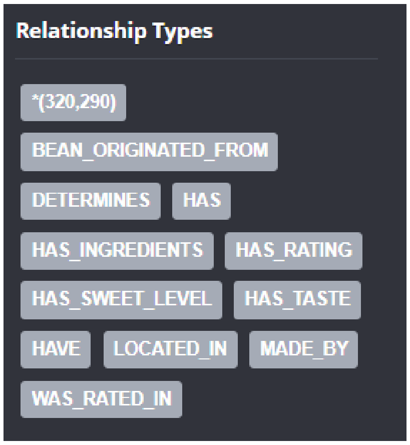
Database Information

Use database

neo4j

Node Labels

*(5,018) Chocolate Company Company_location Country_of_bean_origin Ingredients Province_of_bean_origin Rating Rating_year Sweet_level Taste



Relationship Types

*(320,290) BEAN_ORIGINATED_FROM DETERMINES HAS HAS_INGREDIENTS HAS_RATING HAS_SWEET_LEVEL HAS_TASTE HAVE LOCATED_IN MADE_BY WAS_RATED_IN

Cypher Queries

1. This cypher query looks for companies' name which are in the U.S.A and has a chocolate rating of 4. This information can help me to identify which company has the highest rating that is in the United State, so the shipping cost will be minimized when I purchase their products.

```
MATCH (l:Company_location{name:'U.S.A'})-[r3:LOCATED_IN]-(c:Company)-
[r1:MADE_BY]-(a:Chocolate)-[R2:HAS_RATING]-(r:Rating{rating:'4'})
RETURN DISTINCT c.name AS Copmany,l.name AS Location, r.rating AS Rating
LIMIT 10
```

Result table:

Copmany	Location	Rating
Ruket	U.S.A	4
Sjolinds	U.S.A	4
Meadowlands	U.S.A	4
Mutari	U.S.A	4
Public Chocolatory	U.S.A	4
Exquisito	U.S.A	4
Friis Holm	U.S.A	4
Escazu	U.S.A	4
Argencove	U.S.A	4
Urzi	U.S.A	4

2. This cypher query can determine which company makes the highest rating chocolate flavor and its sweet level.

```
MATCH (c:Company)-[r1:MADE_BY]-(a:Chocolate)-[:HAS_TASTE]-(t:Taste)
MATCH (s:Sweet_level)-[:HAS_SWEET_LEVEL]- (a:Chocolate)-[R2:HAS_RATING]-
(r:Rating{rating:'4'})
RETURN DISTINCT c.name AS Company,t.Taste AS Taste,s.Sweet_level AS Sweet_Level,r.
rating AS Rating LIMIT 10
```

Result table

Company	Taste	Sweet_Level	Rating
Pralus	fourth_taste-lemon	Extra-Bittersweet	4
Pralus	third_taste-ashey	Extra-Bittersweet	4
Pralus	second_taste-burnt	Extra-Bittersweet	4
Pralus	first_taste-creamy	Extra-Bittersweet	4
Frederic Blondeel	first_taste-banana	Bittersweet	4
Shattell	second_taste-perfume	Extra-Bittersweet	4
Shattell	first_taste-strong spice	Extra-Bittersweet	4
Shattell	third_taste-roasty	Extra-Bittersweet	4
Danta	second_taste-floral	Bittersweet	4
Danta	first_taste-sweet spice	Bittersweet	4

3. This cypher query answers my first use case question, and it displayed that cocoa beans from Principe have the highest rating and have the best quality cocoa bean.

```
MATCH (a:Chocolate)-[:BEAN_ORIGINATED_FROM]-(b:Country_of_bean_origin)-[:HAS]-
(d:Province_of_bean_origin)
MATCH (a:Chocolate)-[R2:HAS_RATING]-(r:Rating{rating:'4'})
RETURN DISTINCT b.name AS Bean_origin,d.province AS Province,r.rating AS Rating
LIMIT 10
```

Result table

Bean origin	Province	Rating
Principe	San Juan	4
Principe	Chumphon	4
Principe	O'payo	4
Principe	Tarakan	4
Principe	Philly Blend	4
Principe	Barinas	4
Principe	Talamanca	4
Principe	Presidio	4
Principe	Indianer	4
Principe	Misterio	4

Graph Algorithms

1. Louvain Community Algorithm- These three queries return each country of bean origin name and the ID of the community to which it belongs. In the result tables, it showed 512 countries belong to the same community. This could be that these counties are located near each other.

```
CALL gds.graph.create('origin-related-
entities', ['Company', 'Company_location', 'Country_of_bean_origin', 'Province_of_bean
_origin'], '*')

CALL gds.louvain.stream('origin-related-entities')
YIELD nodeId, communityId
RETURN gds.util.asNode(nodeId).name AS Country, communityId
ORDER BY communityId DESC
```

Partial result table

Country	communityId
Nicaragua	1341
Brazil	1340
Italy	1339
Canada	1338
U.S.A	1337

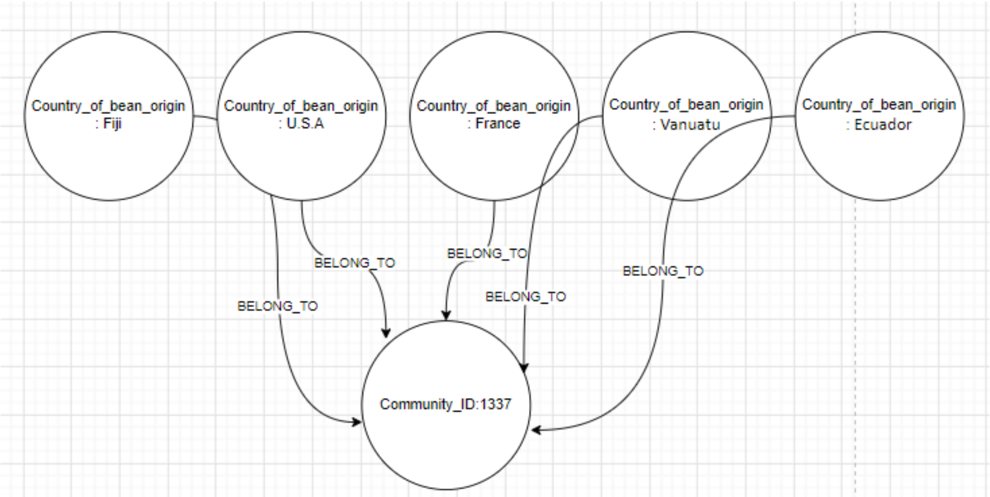
France	1337
Fiji	1337
Vanuatu	1337
Ecuador	1337

```
CALL gds.louvain.stream('origin-related-entities')
YIELD nodeId, communityId
RETURN communityId, COUNT(DISTINCT nodeId) AS members
ORDER BY members DESC
```

Partial result table

communityId	members
1337	512
573	72
574	1
575	1
576	1

Projection Graph


Country_of_bean_origin : Fiji
Country_of_bean_origin : U.S.A
Country_of_bean_origin : France
Country_of_bean_origin : Vanuatu
Country_of_bean_origin : Ecuador
BELONG_TO
BELONG_TO
BELONG_TO
BELONG_TO
BELONG_TO
Community_ID:1337

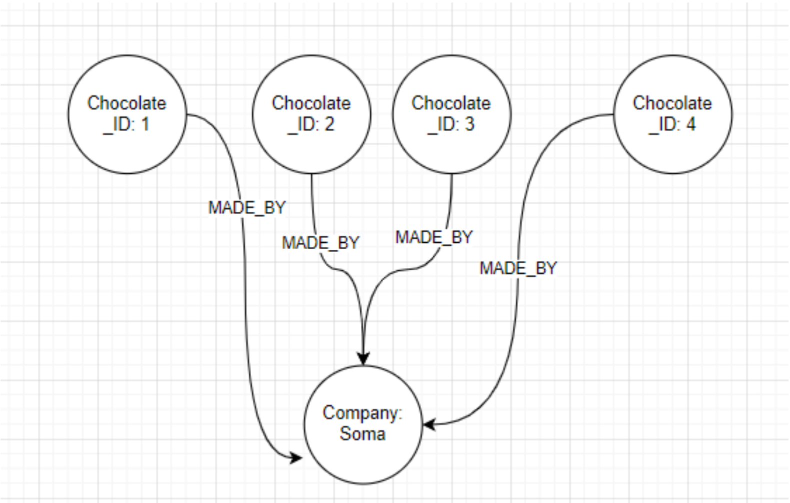
2. PageRank-This query showed the top 10 companies PageRank, and which company made most the chocolate bar, as the result showed 'Soma' have the highest interactions, which makes company 'Soma' the most important node.

```
CALL gds.pageRank.stream('Company-graph') YIELD nodeId, score AS pageRank
WITH gds.util.asNode(nodeId) AS n, pageRank
MATCH (n)-[i:MADE_BY]-()
RETURN n.name AS company, pageRank, count(i) AS interactions
ORDER BY interactions DESC LIMIT 10
```

Result table

company	pageRank	interactions
Soma	0.15	52
Arete	0.15	32
Fresco	0.15	31
Bonnat	0.15	28
Pralus	0.15	26
A. Morin	0.15	25
Valrhona	0.15	22
Domori	0.15	22
Guittard	0.15	22
Zotter	0.15	21

Projection Graph



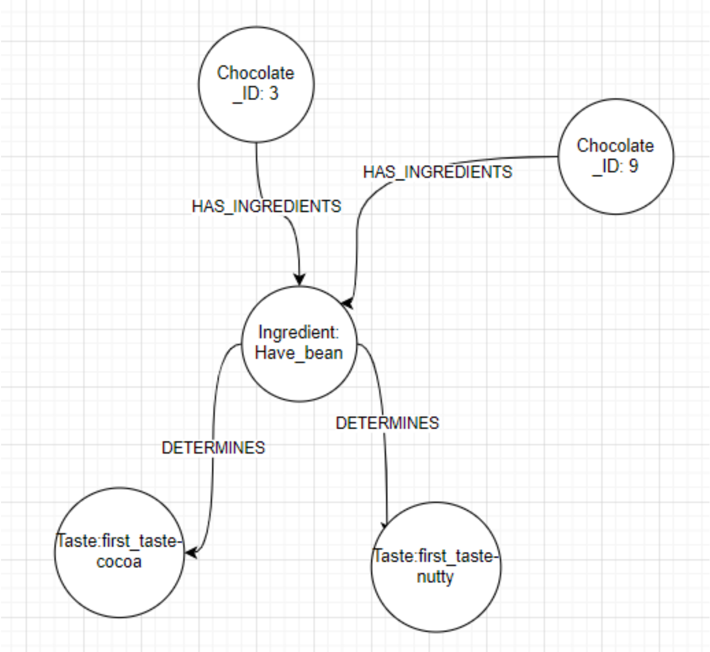
3. Betweenness Centrality- this query finds the node that serves as a bridge from one part of a graph to another. As the result tables show, 'have_bean' has the highest score, which tells me that the cocoa bean is the most important ingredient when making a chocolate bar, and all other tastes including the ingredient of the cocoa bean.

```
CALL gds.graph.create('Chocoate-taste-related-
entities', ['Chocolate', 'Ingredients', 'Taste', 'Sweet_level'], '*')
CALL gds.betweenness.stream('Chocoate-taste-related-entities') YIELD nodeId, score
RETURN gds.util.asNode(nodeId).Ingredients AS Ingredients, score
ORDER BY score DESC LIMIT 10
```

Result table

Ingredients	score
have_bean	429614.7
have_not_salt	422467.6
have_not_sweetener_without_sugar	414933.3
have_sugar	413774.4
have_not_vanila	362391.7
have_not_lecithin	337088.6
have_cocoa_butter	296520.3
have_not_cocoa_butter	133094.4
have_lecithin	92526.14
have_vanila	67223

Projection Graph


Chocolate _ID: 3
Chocolate _ID: 9
HAS_INGREDIENTS
HAS_INGREDIENTS
Ingredient: Have_bean
DETERMINES
DETERMINES
Taste:first_taste-cocoa
Taste:first_taste-nutty