Washington State University
CPT_S 415 – Big Data
Online

Srinivasulu Badri

Assignment 3

Name: Nam Jun Lee

Student Number: 11606459

1. [RDF]

a. Consider a set of natural language sentences collected from Web pages.

i. A human can like another human.
ii. A human can have a gender property of a male, female or non-binary.
iii. A human can be the father of another human.
iv. A human can be the mother of another human.
v. A human can be married to another human.
vi. A human can have a Birth Year property of type “xs:Year”.
vii. If a human is married to another, then they like each other.
viii. If a human is a mother or father, the human is a parent.

Write a RDF schema and give a graphical presentation to describe these relationships.

RDF schema:

```
<rdfs:Class rdf:ID="Human">
    <rdfs:comment>
        The class of Human.
        Superclass of Man, Woman, NonBinary, Year classes
    </rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:ID="AnotherHuman">
    <rdfs:comment>
        The class of AnotherHuman
        This class represents other human
    </rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:ID="Male">
    <rdfs:comment>
        The class of Man
```

```
            Subclass of Human
        </rdfs:comment>
        <rdfs:subClassOf rdf:resource="#Human"/>
</rdfs:Class>

<rdfs:Class rdf:ID="Female">
        <rdfs:comment>
            The class of Female
            Subclass of Human
        </rdfs:comment>
        <rdfs:subClassOf rdf:resource="#Human"/>
</rdfs:Class>
<rdfs:Class rdf:ID="NonBinary">
        <rdfs:comment>
            The class of NonBinary
            Subclass of Human
        </rdfs:comment>
        <rdfs:subClassOf rdf:resource="#Human"/>
</rdfs:Class>

<rdfs:Class rdf:ID="Year">
        <rdfs:comment>
            The class of Year
            Birth Year for Human
            Subclass of Human
        </rdfs:comment>
        <rdfs:subClassOf rdf:resource="#Human"/>
</rdfs:Class>

<rdf:Property rdf:ID="canLike">
```

```
        <rdfs:comment>
            Inherits its domain ("Human") and range ("AnotherHuman")
            from its superproperty "isMarried"
        </rdfs:comment>
        <rdfs:domain rdfs:resource="#Human" />
        <rdfs:range rdfs:resource="#AnotherHuman" />
        <rdfs:subPropertyOf rdfs:resource="#isMarried" />
    </rdf:Property>

    <rdf:Property rdf:ID="canBe">
        <rdfs:domain rdfs:resource="#Human" />
        <rdfs:range rdfs:resource="#Man" />
        <rdfs:range rdfs:resource="#Woman" />
        <rdfs:range rdfs:resource="#NonBinary" />
    </rdf:Property>

    <rdf:Property rdf:ID="isMarried">
        <rdfs:domain rdfs:resource="#Human" />
        <rdfs:range rdfs:resource="#AnotherHuman" />
    </rdf:Property>

    <rdf:Property rdf:ID="isMotherOf">
        <rdfs:comment>
            Inherits its domain ("Woman") and range ("AnotherHuman")
            Inherits its domain ("NonBinary") and range ("AnotherHuman")
            from its superproperty "isParent"
        </rdfs:comment>
        <rdfs:domain rdfs:resource="#Woman" />
        <rdfs:domain rdfs:resource="#NonBinary" />
        <rdfs:range rdfs:resource="#AnotherHuman" />
```

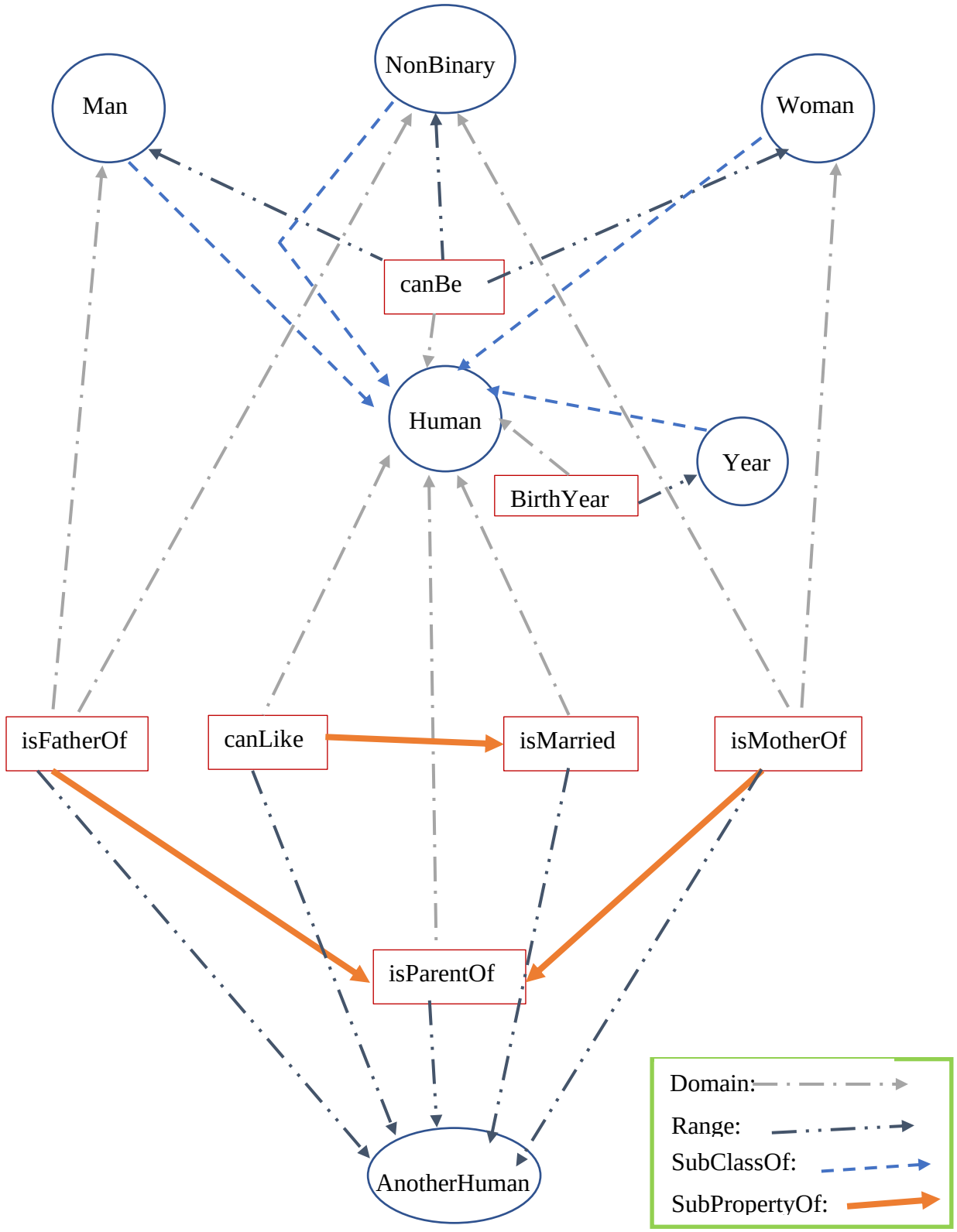
```
        <rdfs:subPropertyOf rdfs:resource="#isParentOf" />
</rdf:Property>

<rdf:Property rdf:ID="isFatherOf">
        <rdfs:comment>
                Inherits its domain ("Man") and range ("AnotherHuman")
                Inherits its domain ("NonBinary") and range ("AnotherHuman")
                from its superproperty "isParent"
        </rdfs:comment>
        <rdfs:domain rdfs:resource="#Man" />
        <rdfs:domain rdfs:resource="#NonBinary" />
        <rdfs:range rdfs:resource="#AnotherHuman" />
        <rdfs:subPropertyOf rdfs:resource="#isParentOf" />
</rdf:Property>

<rdf:Property rdf:ID="isParentOf">
        <rdfs:domain rdfs:resource="#Human" />
        <rdfs:range rdfs:resource="#AnotherHuman" />
</rdf:Property>

<rdf:Property rdf:ID="BirthYear">
        <rdfs:romain rdfs:resource="#Human" />
        <rdfs:range rdfs:resource="#Year" />
</rdf:Property>
```

Graphical Presentation of this schema:



b. Write an instance of the RDF schema in A that express the following with a complete information that can be inferred from the schema.

i. Mary is a woman and she is John's wife.
ii. Sophie, Sandra and Susan are women.
iii. Mary and John has a son Frank.
iv. John was born in 1950.
v. Frank was born in 1980.
vi. Susan is John's daughter.
vii. Susan was born in 1978.
viii. Frank likes Sophie
ix. Sandra likes Frank.

All other information for everyone else is unknown.

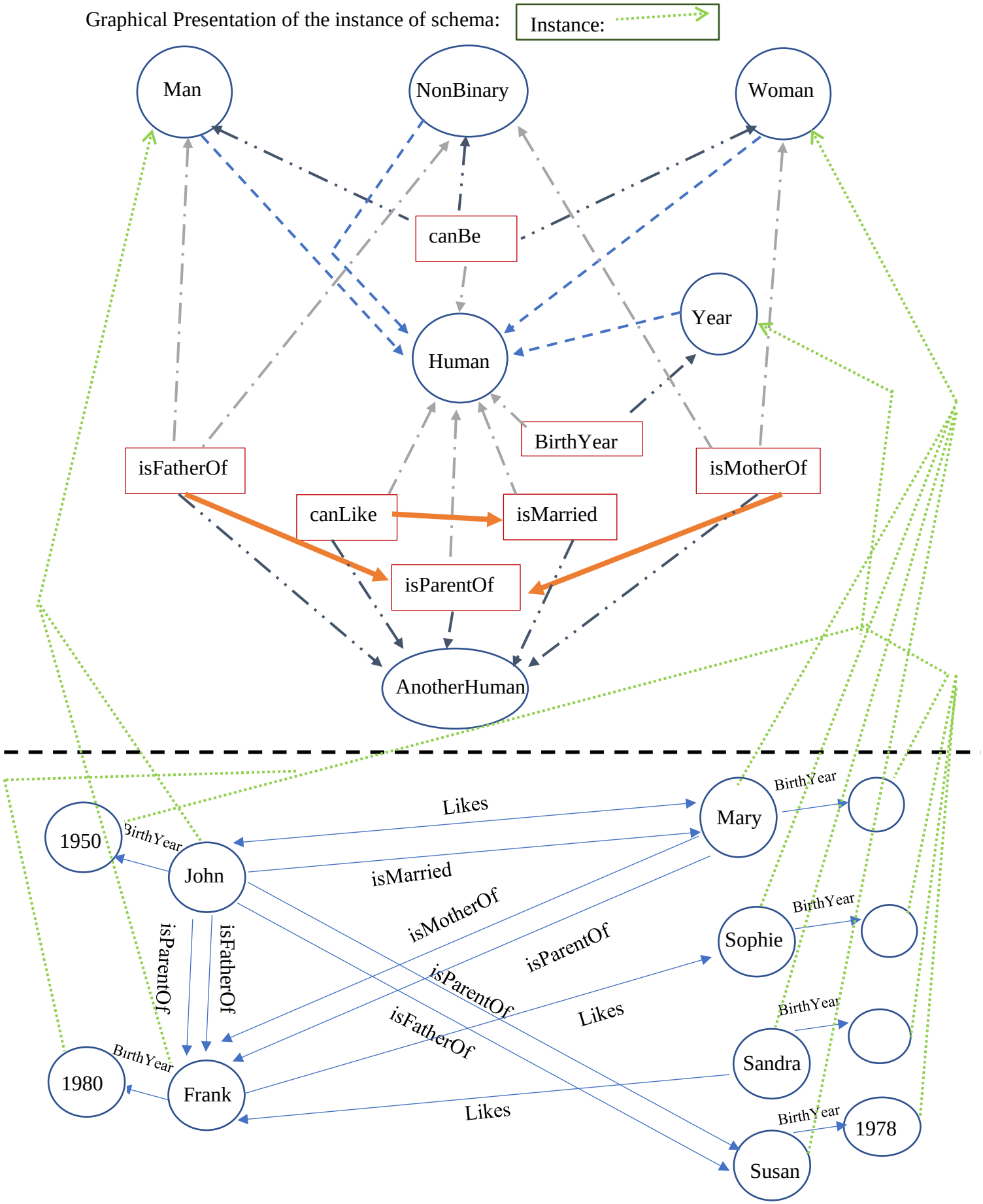
Write a RDF document and give a graphical presentation to describe these facts

RDF document:

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:uni="http://www.mydomain.org/uni-ns">
    <rdf:Description rdf:about="Human1">
        <uni:name>Mary</uni:name>
        <uni:gender>Woman</uni:gender>
        <uni:BirthYear></uni:BirthYear>
        <uni:Likes rdf:resource="Human2"/>
        <uni:isMotherOf rdf:resource="Human3"/>
        <uni:isParentOf rdf:resource="Human3"/>
    </rdf:Description>
    <rdf:Description rdf:about="Human2">
```

```
        <uni:name>John</uni:name>
        <uni:gender>Man</uni:gender>
        <uni:BirthYear>1950</uni:BirthYear>
        <uni:Likes rdf:resource="Human1"/>
        <uni:isMarried rdf:resource="Human1"/>
        <uni:isFatherOf>
            <rdf:Bag>
                <rdf:_1 rdf:resource="Human3"/>
                <rdf:_2 rdf:resource="Human4"/>
            </rdf:Bag>
        </uni:isFatherOf>
        <uni:isParentOf>
            <rdf:Bag>
                <rdf:_1 rdf:resource="Human3"/>
                <rdf:_2 rdf:resource="Human4"/>
            </rdf:Bag>
        </uni:isParentOf>
    </rdf:Description>
    <rdf:Description rdf:about="Human3">
        <uni:name>Frank</uni:name>
        <uni:gender>Man</uni:gender>
        <uni:BirthYear>1980</uni:BirthYear>
        <uni:Likes rdf:resource="Human5"/>
    </rdf:Description>
    <rdf:Description rdf:about="Human4">
        <uni:name>Susan</uni:name>
        <uni:gender>Woman</uni:gender>
        <uni:BirthYear>1978</uni:BirthYear>
    </rdf:Description>
    <rdf:Description rdf:about="Human5">
```

```
        <uni:name>Sophie</uni:name>
        <uni:gender>Woman</uni:gender>
        <uni:BirthYear></uni:BirthYear>
    </rdf:Description>
    <rdf:Description rdf:about="Human6">
        <uni:name>Sandra</uni:name>
        <uni:gender>Woman</uni:gender>
        <uni:BirthYear></uni:BirthYear>
        <uni:Likes rdf:resource="Human3"/>
    </rdf:Description>
</rdf:RDF>
```


Graphical Presentation of the instance of schema:
Instance:
Man
NonBinary
Woman
canBe
Year
Human
BirthYear
isFatherOf
isMotherOf
canLike
isMarried
isParentOf
AnotherHuman
1950
BirthYear
John
Likes
Mary
BirthYear
isMarried
isMotherOf
isParentOf
isFatherOf
isParentOf
isParentOf
Sophie
BirthYear
Likes
isFatherOf
BirthYear
1980
Frank
Sandra
Likes
BirthYear
1978
Susan

3. [Property Graph]

Create a Labeled Property Graph to describe the project management data shown in Lecture 2-3. Describe the vertices and the edges along with their labels and properties. Draw the graph using a Graph database tool (Neo4j or other)

Note: Refer to the table included in the attachment with this assignment.

This data has a total of seven labels:

1. Project (4)

2. Employee (14)

3. Assign_hours (21)

4. Job (9)

5. Participated_In (21)

6. Worked_for (21)

7. Job_Is. (14)

1 to 4 labels are vertices, and 5 to 7 labels are edges.

The properties of Each label:

Project: PROJ_NUM, PROJ_NAME

Employee: EMP_NUM, EMP_NAME, EMP_HIREDATE

Job: JOB_CODE, JOB_CLASS, CHR_HOUR

Assign: PROJ_NUM, EMP_NUM, ASSIGN_HOURS

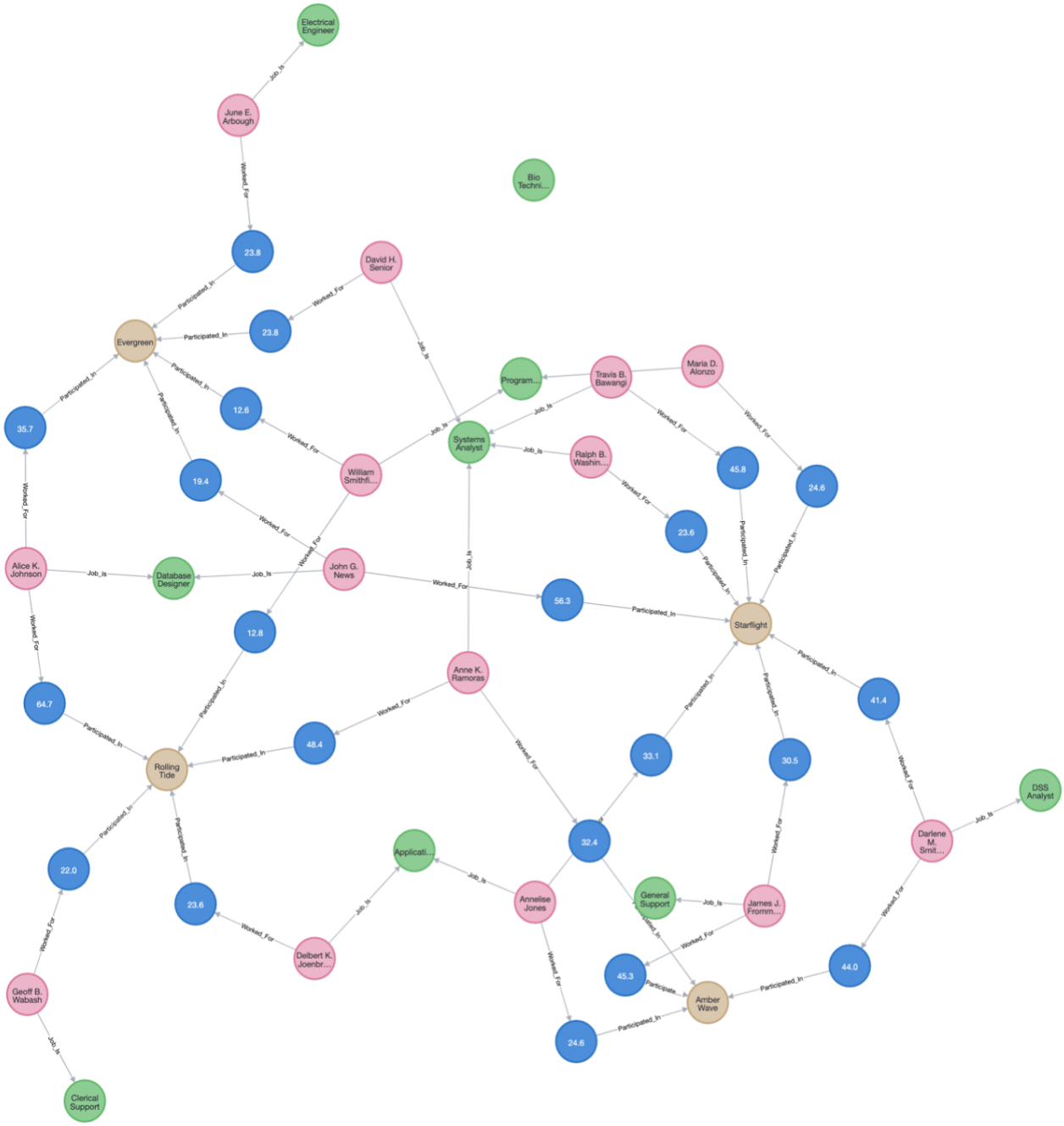
Connect Information:

Participated_In edges connect Project vertices and Assign_hours vertices.

Worked_For edges connect Employee vertices and Assign_hours vertices.

Job_Is edges connect Employee vertices and Job vertices.

Graph using Neo4j:



4. **[Approximate Query Processing]**

This question continues our discussion on using data synopsis for query processing based on data- driven approximation. You are given a vector of numbers: [127, 71, 87, 31, 59, 3, 43, 99, 100, 42, 0, 58, 30, 88, 72, 130], each data point records the frequency of communication of a server in a 5-minute interval. For example, in the first 5 minutes, 127 contacts are observed. In the next 5 minutes, 71 contacts, ...

Note: For these questions, similar to the examples shown in the lectures, discard the lowest level (high-resolution) coefficients (i.e. only keep the first 50% of coefficients).

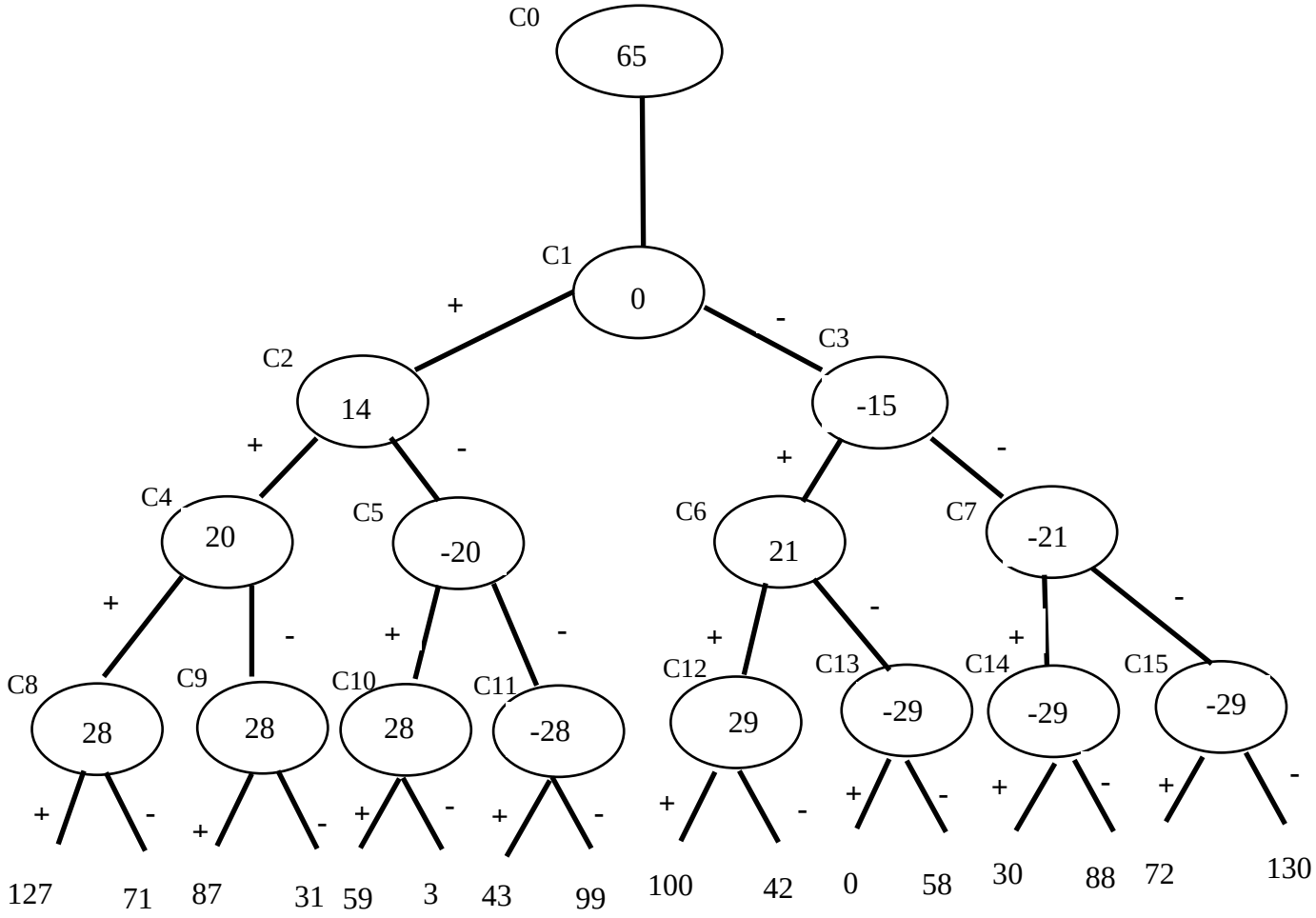
Resolution	Averages	Detail Coefficients
4	[127, 71, 87, 31, 59, 3, 43, 99, 100, 42, 0, 58, 30, 88, 72, 130]	----
3	[99, 59, 31, 71, 71, 29, 59, 101]	[28,28,28,-28,29,-29,-29,-29]
2	[79, 51, 50, 80]	[20,-21,21,-21]
1	[65, 65]	[14,-15]
0	[65]	[0]

a. Give the Haar decomposition and draw a corresponding error tree for the contacts data vector.

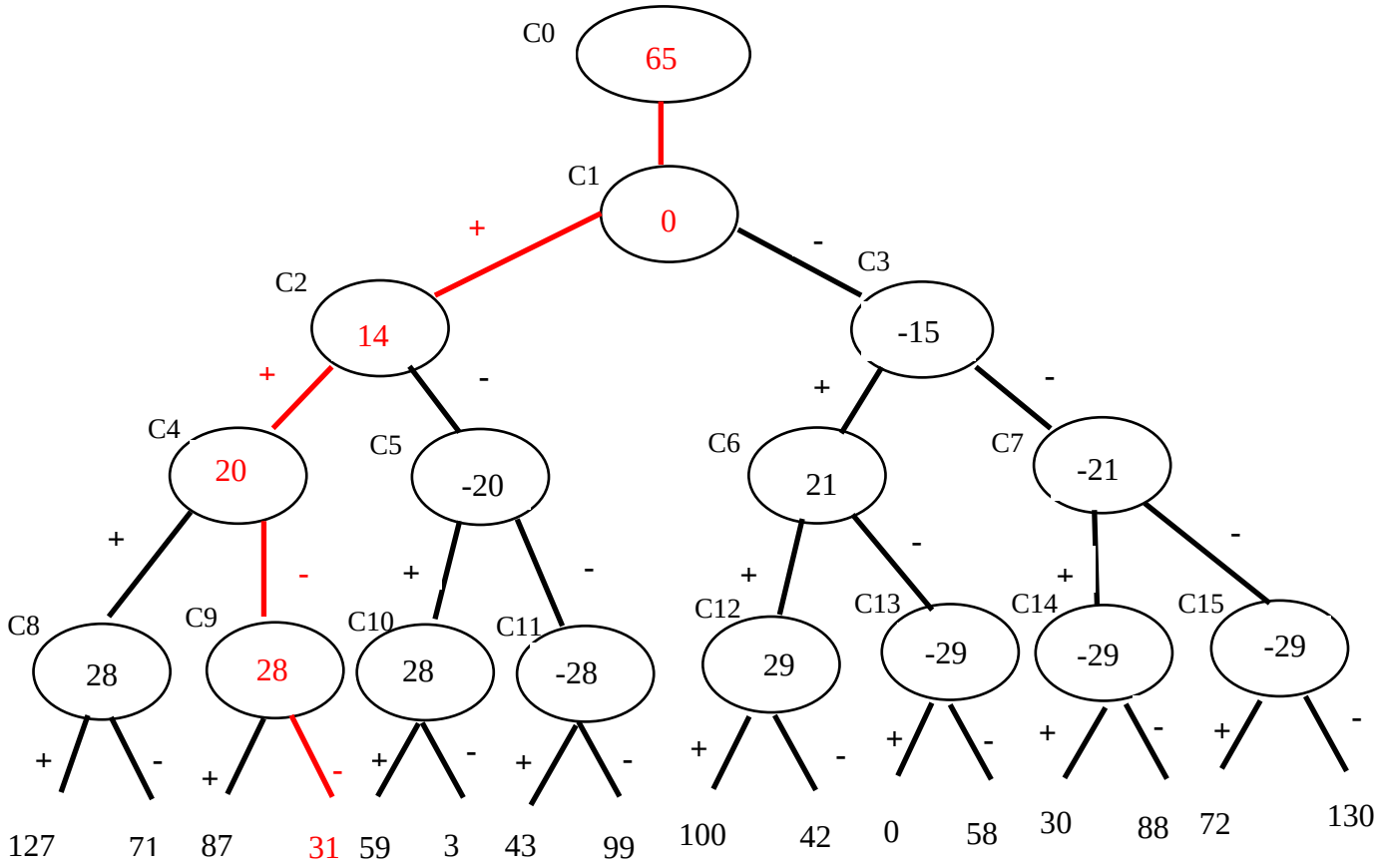
According to table above,

Haar decomposition: [65, 0, 14, -15, 20, -21, 21, -21, 28, 28, 28, -28, 29, -29, -29, -29]

Error tree:


C0
65
C1
0
+
-
C2
14
C3
-15
+
-
+
-
C4
20
C5
-20
C6
21
C7
-21
+
-
+
-
+
-
+
-
C8
28
C9
28
C10
28
C11
-28
C12
29
C13
-29
C14
-29
C15
-29
+
-
+
-
+
-
+
-
+
-
+
-
+
-
127
71
87
31
59
3
43
99
100
42
0
58
30
88
72
130

b. Give the process and result for reconstructing the frequency during time interval [15, 20] using Haar decomposition



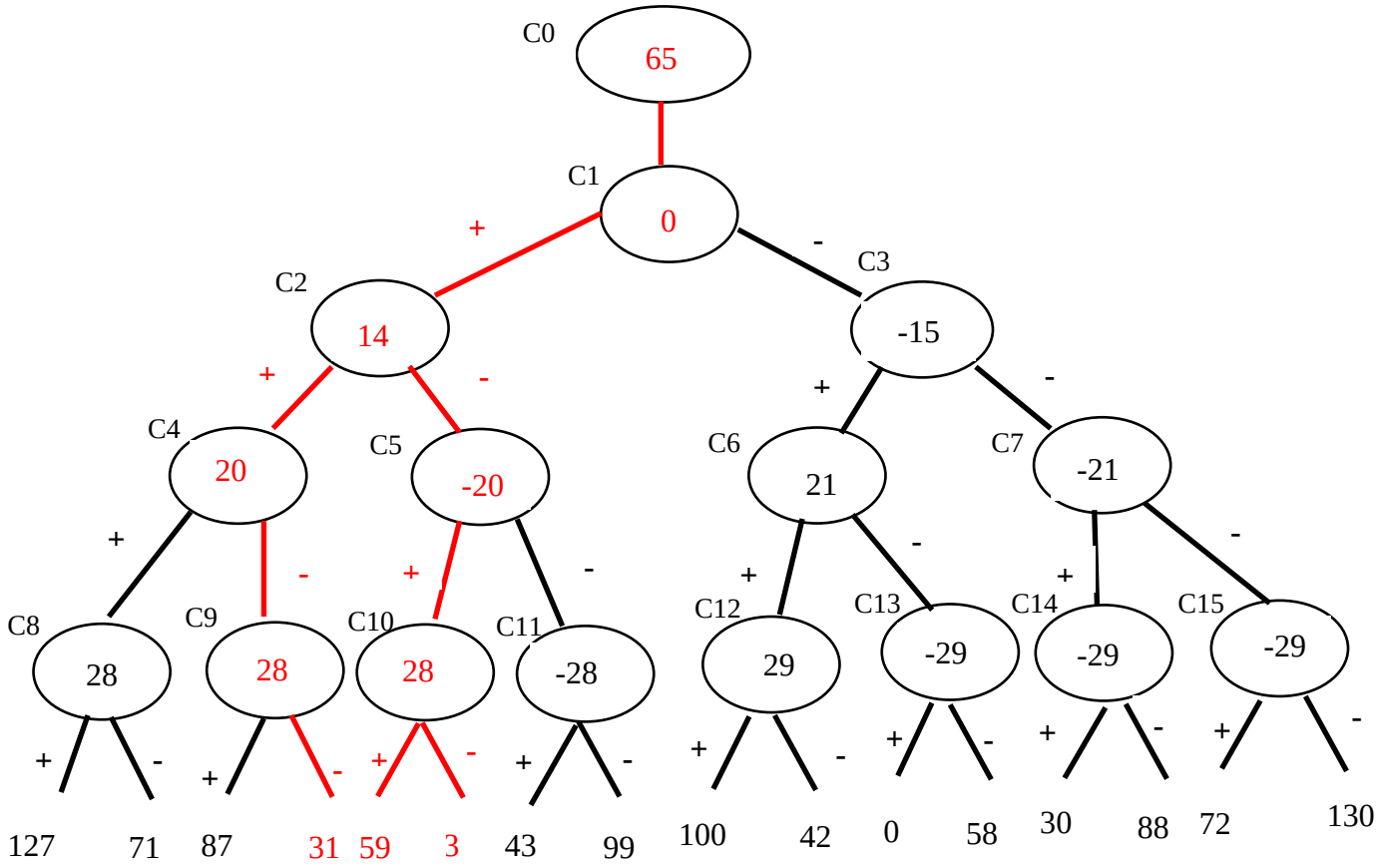
The frequency during time interval [15,20] is shown in red below. Where C0, C1, C2, C4, C9 are coefficients.

Therefore, A[15,20] = C0 + C1 + C2 – C4 – C9

= 65 + 0 + 14 – 20 – 28

= 31

c. Use Haar decomposition and error tree to compute the total number of communications between time interval [15, 30].



To calculate the total number of communications [15, 30] between time intervals, we must first find the time intervals [15, 20], time intervals [20, 25], and time intervals [25, 30].

The frequency during time interval [15,20] is shown in red below. Where C0, C1, C2, C4, C9 are coefficients.

Therefore, A[15,20] = C0 + C1 + C2 – C4 – C9

= 65 + 0 + 14 – 20 – 28

= 31

The frequency during time interval [20,25] is shown in red below. Where C0, C1, C2, C5, C10 are coefficients.

Therefore, A[20,25] = C0 + C1 - C2 + C5 + C9

= 65 + 0 - 14 + (-20) + 28

= 59

The frequency during time interval [25,30] is shown in red below. Where C0, C1, C2, C5, C10 are coefficients.

Therefore, A[25,30] = C0 + C1 - C2 + C5 - C9

= 65 + 0 - 14 + (-20) - 28

= 3

Hence, the total number of communications between time interval [15, 30],

A[15,30] = A[15,20] + A[20,25] + A[25,30]

= 31 + 59 + 3

= 93