АНАЛИЗ ГРАФОВОЙ БАЗЫ ДАННЫХ

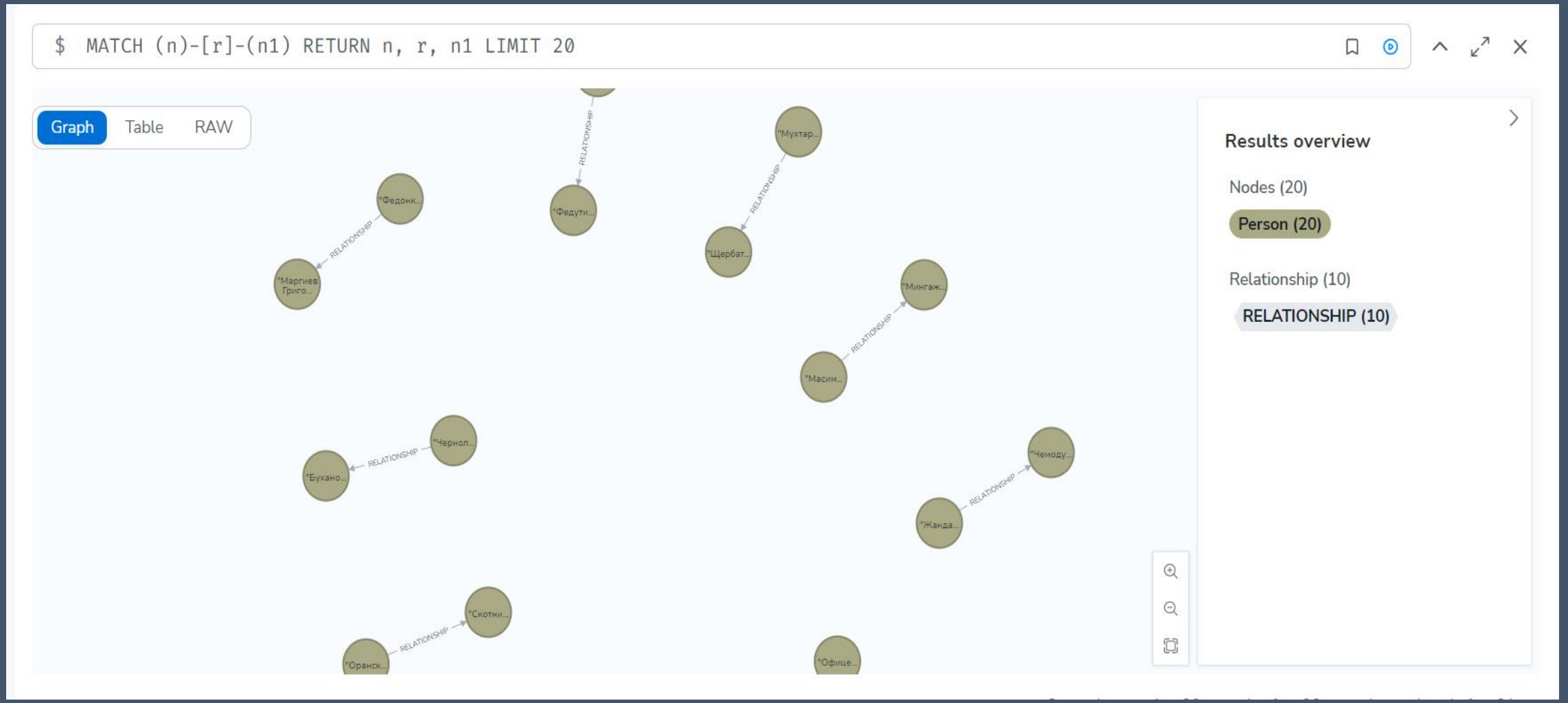
на примере **neo4j**

Агалян Роберт

Библиотеки для работы и анализа графовой базы данных в Python

- Numpy
- Pandas
- Requests
- Io
- Neo4j
- Py2neo
- Networkx
- Matplotlib
- pyvis

Neo4j позволяет загружать данные, получить визуальное представление бд и реализовывать запросы к ней не прямо во встроенном neo4j workspace.



Создав подключение к neo4j и загрузив бд из гугл диска в переменную **data_**, код ниже загрузит данные переменной в neo4j.

```
with driver.session() as session:
    query = """
        UNWIND $data_ AS row
        MERGE (n1:Person {name: row.name1})
        MERGE (n2:Person {name: row.name2})
        MERGE (n1)-[r:RELATIONSHIP {id: row.id}]-(n2)
            """
    session.run(query, data_=data_)
```

Проведя предварительный анализ, мы узнали:

- Большинство узлов имеют одну связь
- Остальные узлы имеют больше одной связи
- Среди них есть узлы с количеством связей до 50.

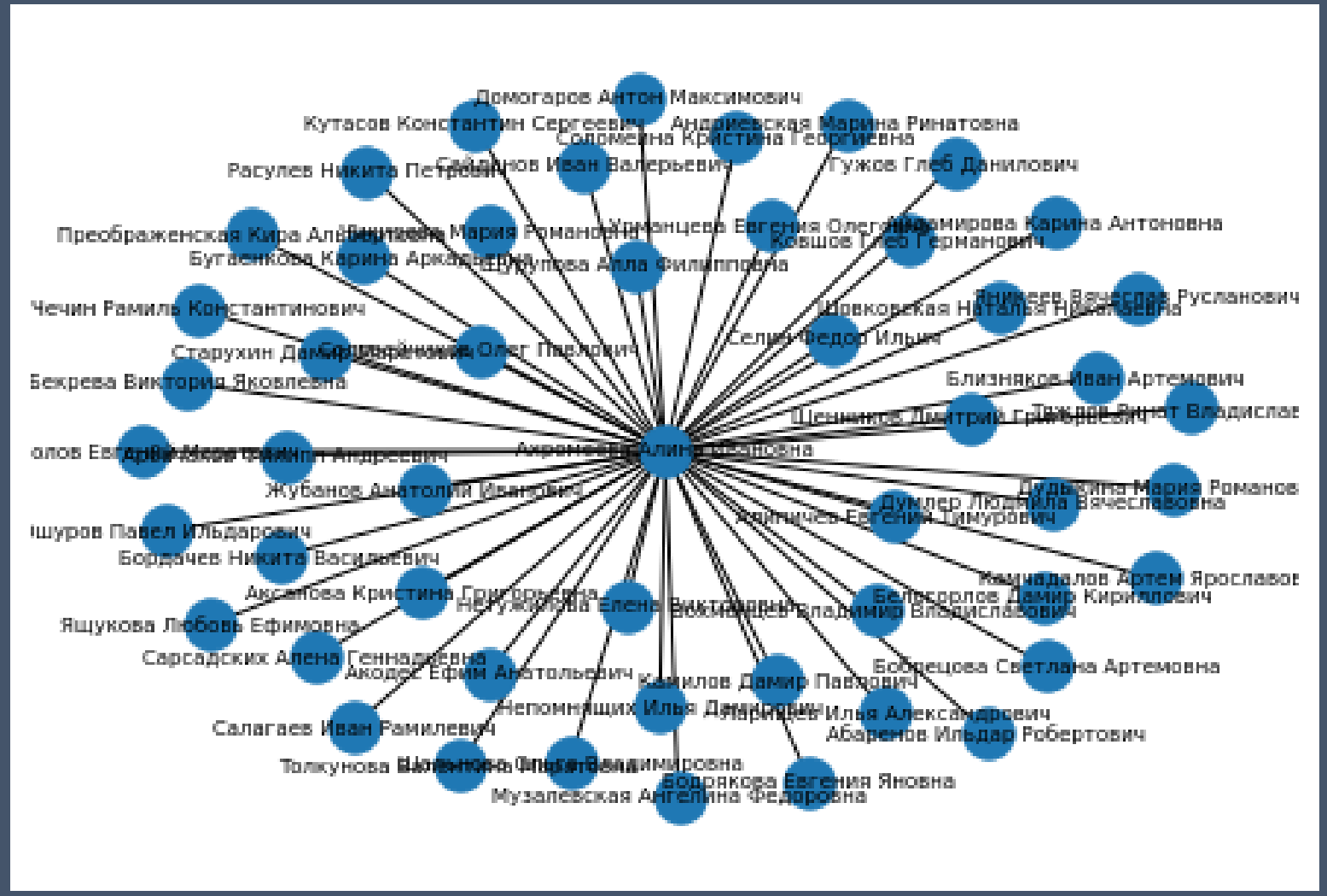
Мы пронализировали крупнейшие по количеству связей узлы.

Проведя предварительный анализ, мы узнали:

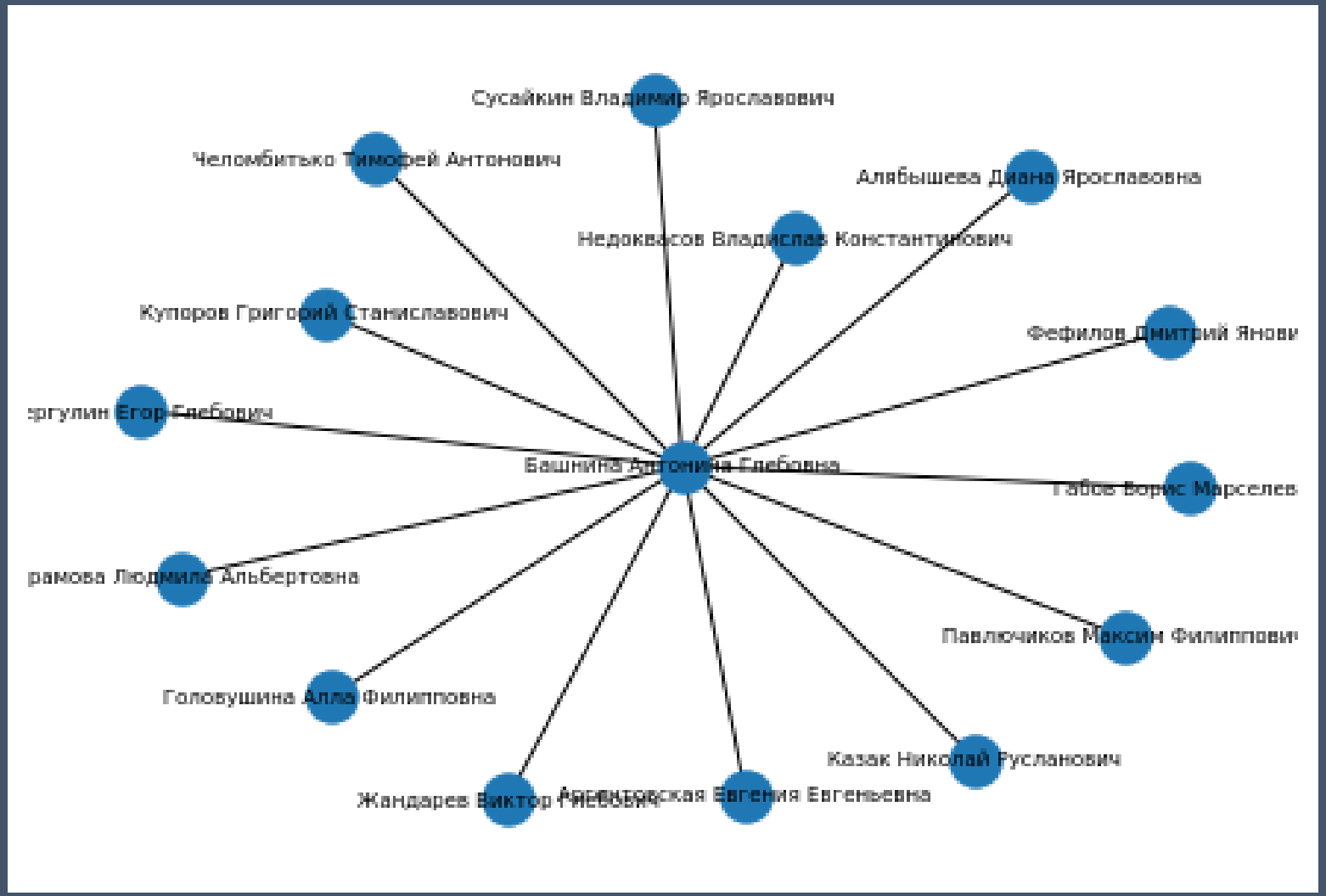
- Большинство узлов имеют одну связь
- Остальные узлы имеют больше одной связи
- Среди них есть узлы с количеством связей до 50.

Мы пронализировали крупнейшие по количеству связей узлы.

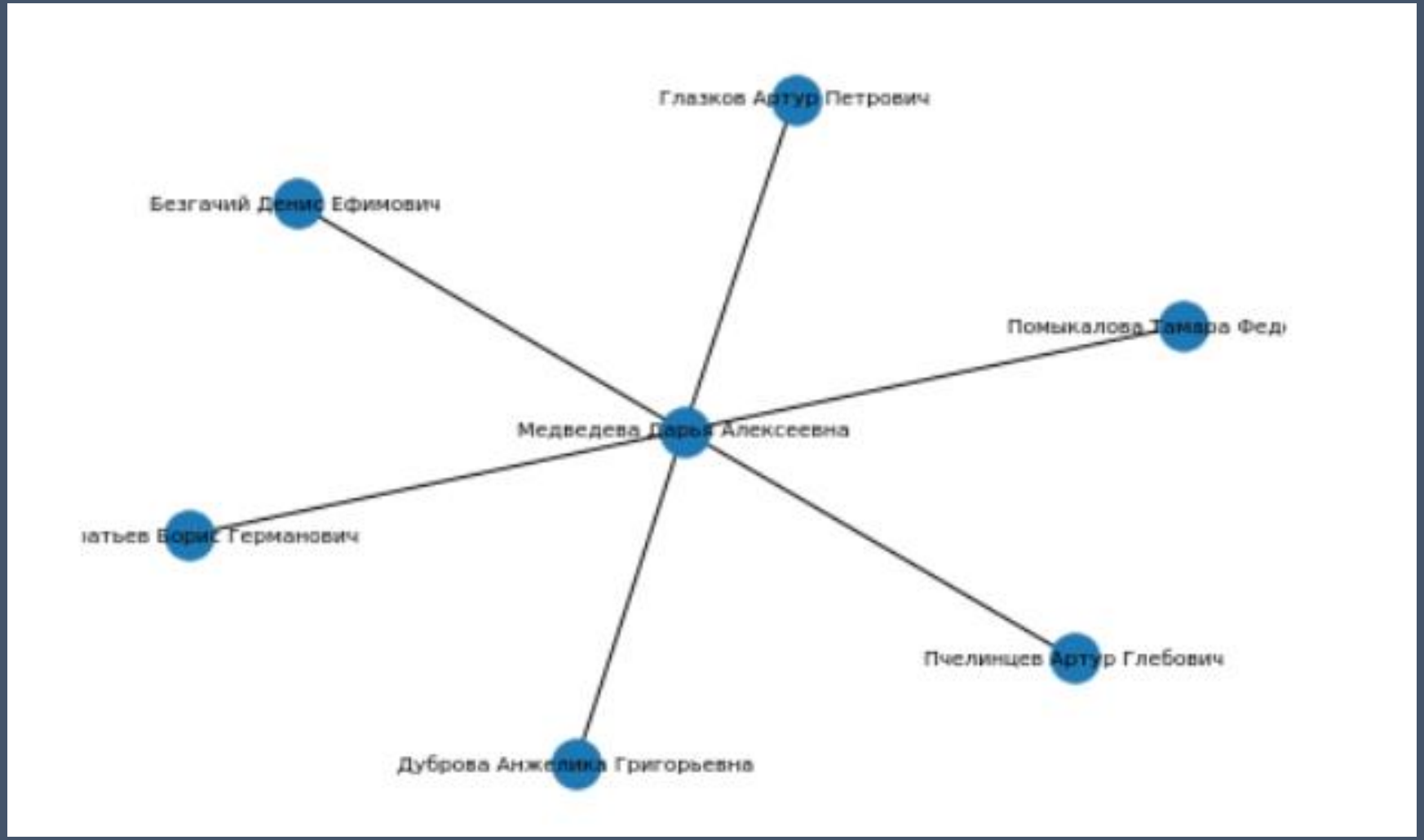
Узел с именем «Ахромеева Алина Ивановна» имеет 50 связей, это целое сообщество.



Другим менее большим сообществом можно назвать все узлы связанные с узлом «Башнина Антонина Глебовна»



Третий подобный потенциальный центр сообщества - узел «Медведева Дарья Алексеевна»



Среди всех узлов есть лишь один узел, связанный с двумя разными узлами различными связями.

```
[{'n1': Node('Person', name='Шолохов Игорь Робертович'),
  'n2': Node('Person', name='Ляуданский Валентин Владиславович'),
  'commonNodes': [Node('Person', name='Торгунаков Роман Кириллович'),
   Node('Person', name='Пафомова Кира Вадимовна')]},
 {'n1': Node('Person', name='Пафомова Кира Вадимовна'),
  'n2': Node('Person', name='Торгунаков Роман Кириллович'),
  'commonNodes': [Node('Person', name='Ляуданский Валентин Владиславови
ч'),
   Node('Person', name='Шолохов Игорь Робертович')]}]
```

Используя библиотеки python мы узнали:

1. 5 узлов имеющих наибольшее количество связей не имеют общих связанных узлов, то есть сообщества не пересекаются.
2. Более того, среди всех узлов, имеющих более трёх связей нет пересечений.
3. Среди узлов, имеющих меньшее количество связей есть пересечение, всего одно.

Меры центральности

Используя библиотеку для сетевого анализа **networkx** исследуем бд:

Рассчитаем две меры центральности: центральность по степени и центральность по собственному вектору в качестве мер влияния узлов бд.

Распределение значений центральности по степени выглядит примерно так

```
collections.Counter(degree_centrality.values())
```

```
Counter({0.0001010305112143867 4: 9872,
         0.0050515255607193375: 1,
         0.0014144271570014143: 1,
         0.0006061830672863204: 1,
         0.0003030915336431602: 2,
         0.0002020610224287734 8: 19,
         0.0005051525560719338: 2,
         0.0004041220448575469 6: 1})
```

Подобное распределение объяснимо нашими предыдущими выводами о количестве связей у узлов бд.

Распределение значений центральности по собственному вектору выглядит слегка иначе, это объяснимо тем, что расчёт происходит иначе

```
collections.Counter(eigenvector_centrality.values())
```

```
Counter({1.2475527207619724e-16: 9794,
         0.7075373463769591: 1,
         0.09993907178685205: 50,
         6.973707596199351e-07: 1,
         1.8637993229981112e-07: 14,
         7.274944765066784e-11: 6,
         1.781990258712493e-10: 1,
         3.013934270411315e-10: 1,
         2.0833427498170872e-10: 1,
         8.311014201971948e-11: 2,
         3.52416460225304e-11: 8,
         3.1502163707650984e-12: 6,
         3.87743370338239e-10: 1,
         2.492803636003745e-10: 1,
         5.5485947023920616e-11: 1,
         7.075665040031119e-11: 2,
         5.503155999136639e-11: 2,
         1.2515772974232443e-10: 2,
         2.101451184031284e-10: 1,
         1.5078600731945607e-10: 1,
         4.6670576329453154e-11: 1,
         1.7638785570319173e-10: 1,
         2.326017588478472e-10: 1})
```

Наибольшие значение обеих метрик принадлежат узлу «Ахромеева Алина Ивановна»

Исходя из проведённого анализа мы можем сделать следующие выводы:

1. Большинство узлов в данных имеют одну единственную связь.
2. В базе есть три крупных сообщества с кол-вом связей>5.
3. Практически все сообщества и узлы не имеют пересекающихся узлов (которые были бы связаны с другими узлами).
3. Самыми влиятельными является узлы с именами «Ахромеева Алина Ивановна» и «Башнина Антонина Глебовна».

Спасибо за уделённое время