

# Анализ учебных планов и РПД

# Постановка задачи

1. Выстроить граф для всех учебных планов всех образовательных программ реализуемых в университете по программам бакалавриата/специалитета в 2019-2020 учебном году;
2. Провести анализ выстроенного графа;
3. Провести кластерный анализ по содержанию рабочих программ дисциплин реализуемых в университете.

# Граф учебных планов

На основе собираемых от преподавателей (авторов РПД) данных через информационную систему для автоматизации управления учебным процессом разработанную лабораторией ММИС, а именно данных по пререквизитам и постреквизитам из каждой РПД необходимо выстроить граф взаимосвязанных между собой дисциплин (или невзаимосвязанных).

# Анализ графа

**Выстроенный граф по каждому учебному плану каждой образовательной программы необходимо проанализировать на предмет:**

- Логической связности дисциплин имеющих связь в графовой модели;
- Логической последовательности изучения дисциплин;
- Выявление наиболее «важных» дисциплин используя алгоритм позволяющий измерить транзитивное влияние и связность (PageRank) всех узлов (дисциплин) графа, разработанный Google для ранжирования веб-страниц.

# Кластерный анализ

Кластерный анализ по содержанию РПД позволяет выявить «клонов» среди накопленного годами массива текстовых данных РПД. По сути, задача найти схожие по содержанию РПД и объединить их в кластеры.

Методы поиска схожих документов:

- MinHash – алгоритм для расчета показателя схожести двух множеств использующий в качестве оценки коэффициент Жаккара. Применяется в программах «антиплагиат»;
- Word2vec – технология разработанная Google, для анализа семантики естественных языков, основанная на дистрибутивной семантике, машинном обучении и векторном представлении слов. Вкупе с национальным корпусом русского языка который содержит в том числе векторы отражающие семантическую близость между словами, данная технология позволяет провести «глубокий» семантический анализ текстов.

# Этапы

## 1. Получение данных из базы данных лаборатории ММИС:

- Ознакомление со схемой реляционной базы данных лаборатории;
- Написание SQL запроса на выборку необходимых данных;
- Выполнение запроса и экспорт данных.

## 2. Предварительный анализ полученных данных посредством python библиотек pandas и нимру:

- Нахождение аномалий, ошибок в запросе, в данных;
- Анализ ошибок в данных, выявление причин их появления;
- Разделение выгруженных данных на несколько таблиц необходимых для импорта в графовую модель и последующего кластерного анализа.

## 3. Построение графа по выгруженным данным:

- Написание запроса на импорт данных из реляционной модели в графовую. Используя язык запросов Сургер и графовую СУБД Neo4j.

# Этапы

## 4. Анализ полученной графовой модели данных:

- Написание запроса на языке Cypher на выборку дисциплин имеющих связь по типу «пререквизит» в старшем семестре;
- Визуализация запроса посредством Neo4j Browser;
- Запись результата запроса в таблицу;
- Написание запроса рассчитывающего транзитивное влияние и связность (PageRank) всех узлов (дисциплин) графа;
- Подбор инструмента визуализации графа с учетом рассчитанного коэффициента PageRank;
- Визуализация графа с учетом коэффициента PageRank для каждого узла посредством инструмента Graph Data Science Playground.

# Этапы

## 5. Кластерный анализ РПД:

- Изучение структуры РПД;
- Выявление наиболее значимых блоков РПД;
- Определение коэффициента значимости блока;
- Приведение выгруженных и разделенных для кластерного анализа данных к виду, необходимому для алгоритма MinHash и Word2vec;
- Модификация алгоритма, оптимизация под обработку больших объёмов текстовой информации;
- Поблочный расчет схожести РПД попарным сравнением (каждый с каждым) модифицированным алгоритмом MinHash;
- Кластеризация используя технологию Word2vec;
- Запись результатов;
- Импорт результатов в построенную ранее графовую модель с целью визуального анализа;
- На основе визуального анализа определение гиперпараметров:
  - степень схожести (в %) для отнесения РПД к одному кластеру (для MinHash);
  - коэффициент значимости блока РПД.
- Визуализация и запись результатов кластерного анализа
- Анализ построенной графовой модели на основе выходных данных алгоритма MinHash и технологии word2vec.

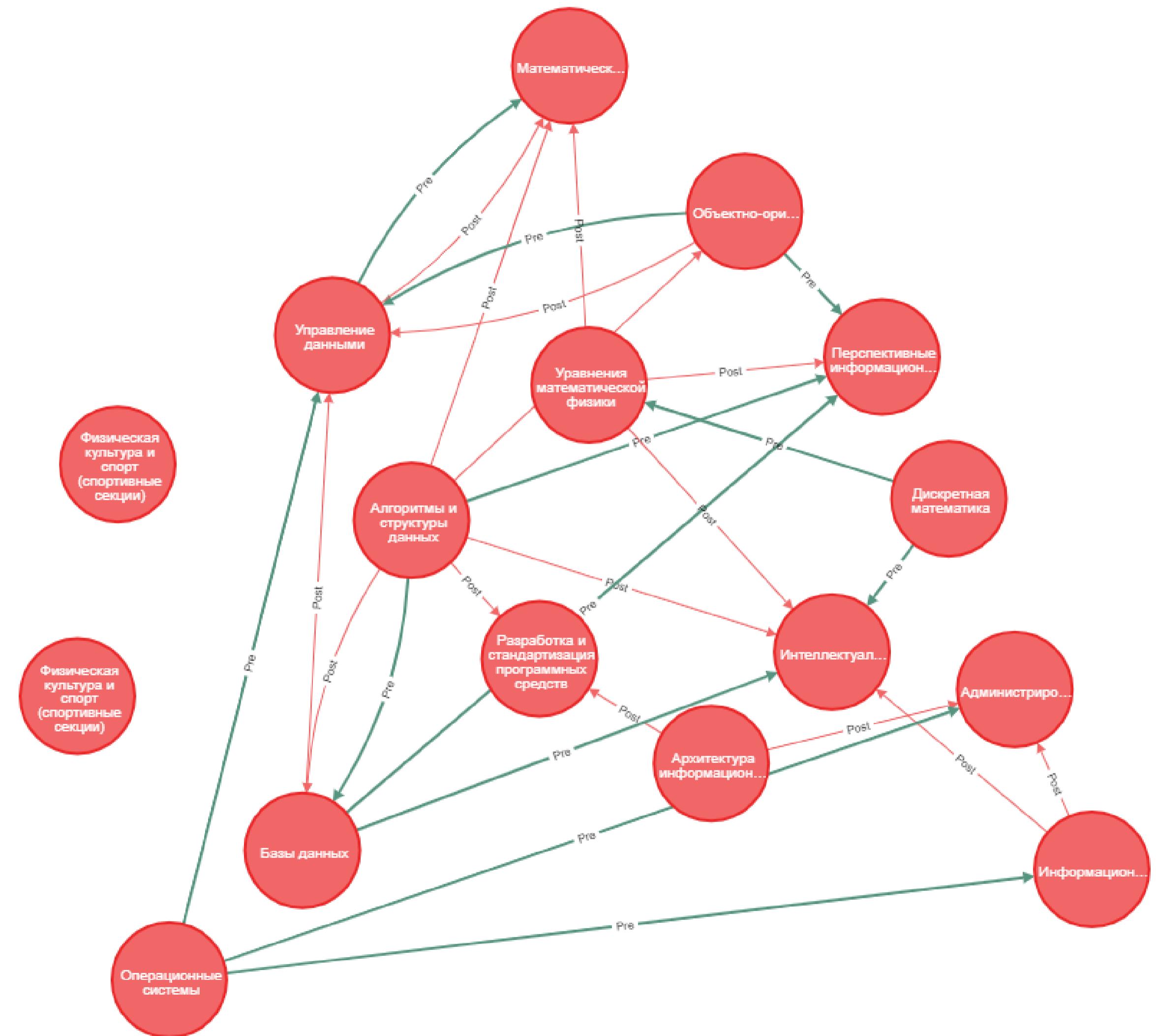
# Исходные данные

В ходе выгрузки с базы данных лаборатории ММИС было получено:

- **33568 записей** содержащих информацию РПД по дисциплинам бакалавров очной формы обучения реализуемых в 2019-2020 учебном году;
- **639090 записей** содержащих связи между дисциплинами по пост и пререквизитам суммарно, между дисциплинами бакалавров очной формы обучения реализуемых в 2019-2020 учебном году.

Рисунок 1 – Визуализация выборки дисциплин учебного плана «B090302ВИС\_09\_4-19plx», дисциплин имеющих код цикла «Б1.Б».

На рисунке видны связи зеленого и красного цветов. Это связи построенные по полу пререквизитов и по полу постреквизитов соответственно. Которые не согласуются на уровне информационной системы лаб. ММИС. Мы видим что например дисциплина «Базы данных» ссылается в качестве пререквизита на дисциплину «Алгоритмы и структуры данных». А дисциплина «Алгоритмы и структуры данных» ссылается на «Базы данных» в качестве постреквизита. Это пример согласованных связей. А например «Управление данными» ссылается в качестве пререквизита на «Операционные системы», но «Операционные системы», не ссылаются на «Управление данными» в качестве постреквизита. Такая малая выборка была выбрана для осязаемости результата



# Результаты

В ходе предварительного анализа выгруженных из БД лаборатории ММИС данных была выявлена архитектурная ошибка, а именно: для каждой отдельно взятой РПД преподаватель (автор РПД) имеет возможность указать как пререквизиты так и постреквизиты, при этом в информационной системе не реализован механизм позволяющий контролировать согласованность между связями. Отсутствие такого механизма позволяет провести дополнительный анализ «органической» согласованности связей по пререквизитам и постреквизитам. Визуализация данной несогласованности представлена на рисунке 1.

# Анализ несогласованности связей

Из рисунка 1 видно что доля несогласованных связей велика на малой выборке. С целью выяснения конкретного процента таких несогласованных связей был написан алгоритм на языке python используя библиотеку numba. Данная библиотека позволила эффективно интерпретировать функцию подсчета количества согласованных связей с высокого уровня абстракции языка python до уровня абстракции языка fortran, что в конечном итоге позволило выполнить за 10 секунд порядка 20 млрд сравнений чисел типа integer. Для сравнения прогнозируемое время выполнения этой же операции без использования библиотеки numba было порядка 18 тыс. секунд, что в 1800 раз быстрее.

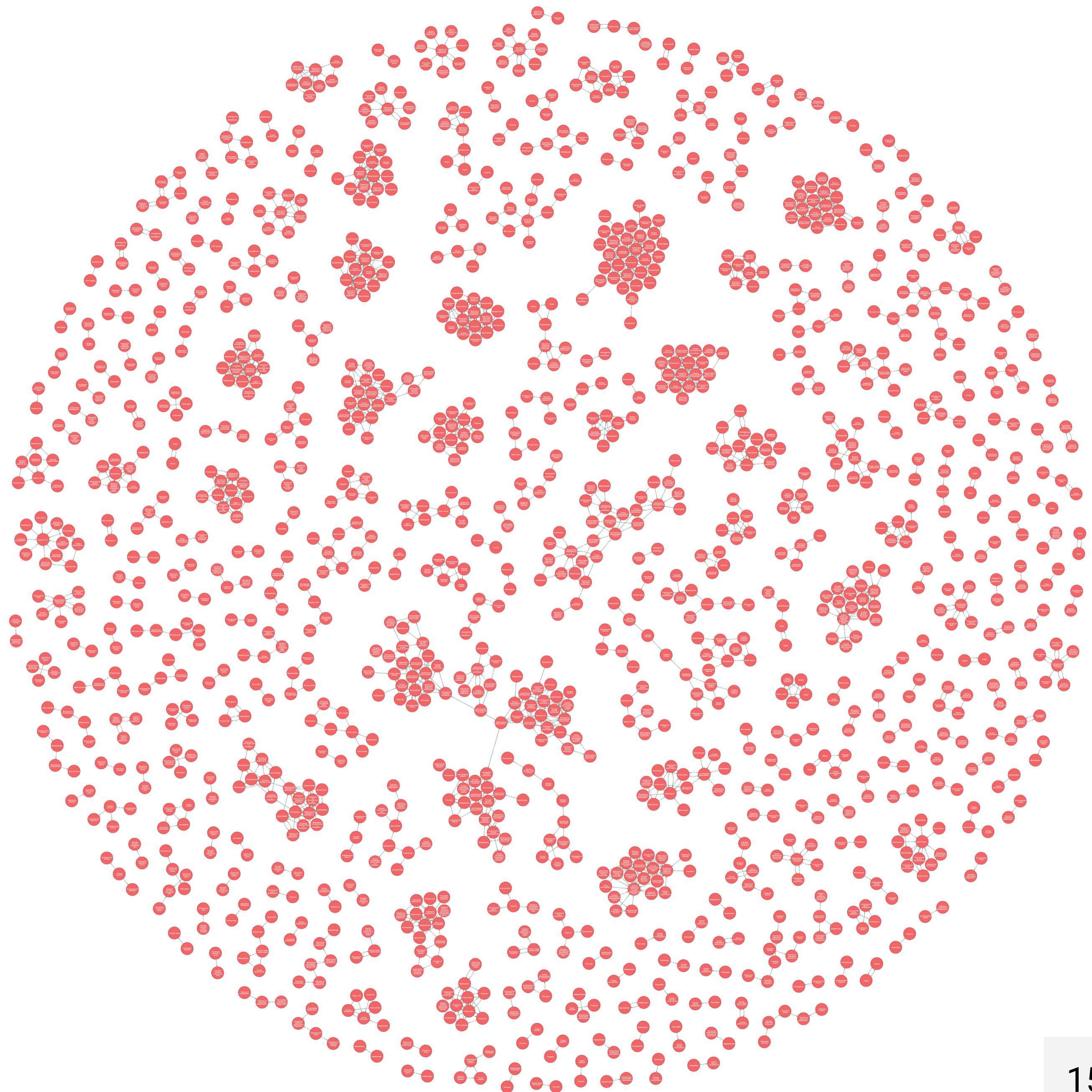
# Результаты

- Согласованными по отношению к связям по типу пререквизит оказалось **28% связей**;
- Согласованными по отношению к связям по типу постреквизит оказалось **30% связей**.

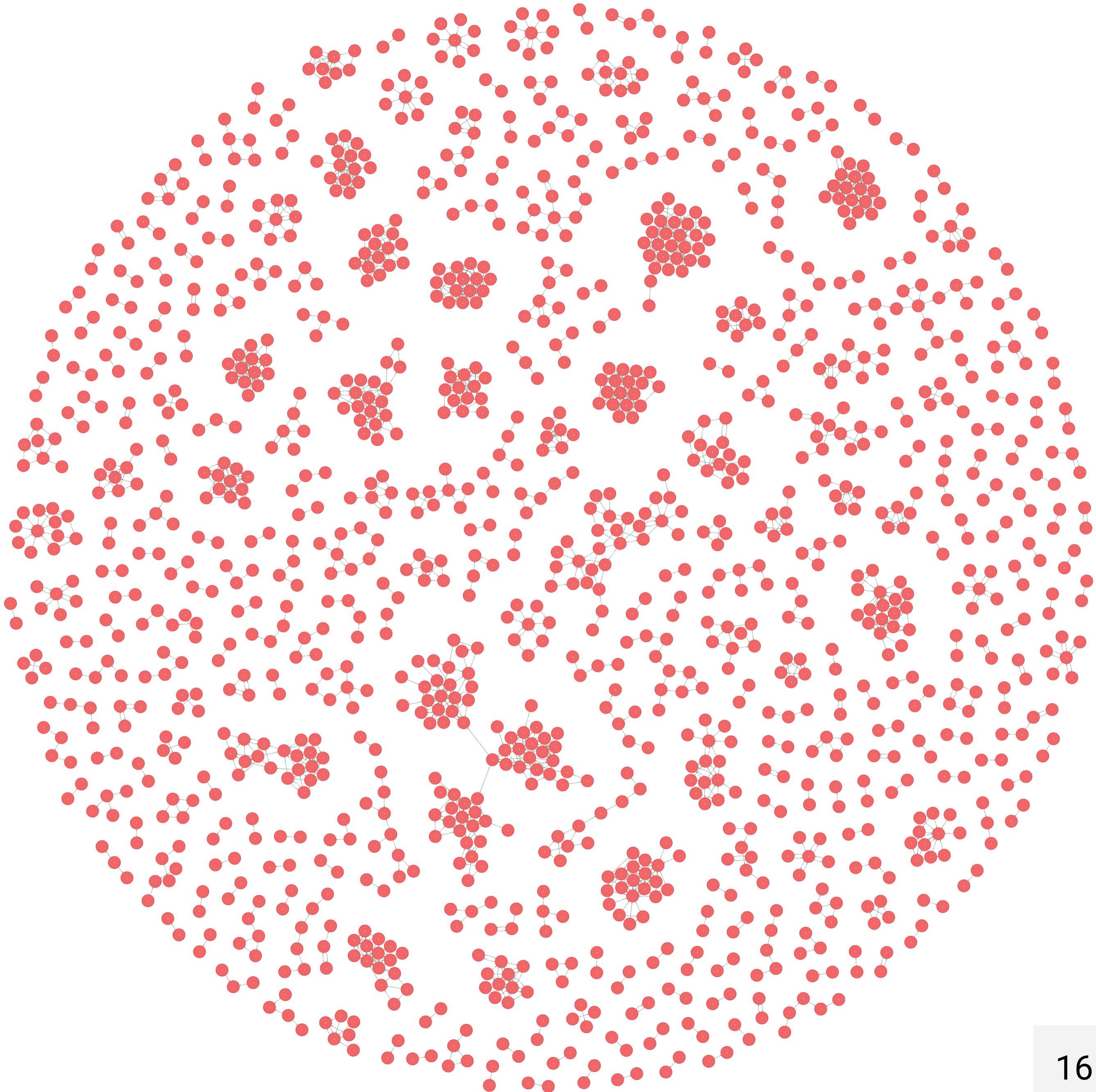
# Анализ логической последовательности изучения дисциплин

В ходе анализа логической последовательности изучения дисциплин был написан запрос к сформированной графовой базе данных, позволяющий посчитать количество связей в качестве пререквизитов на дисциплины стоящие в старшем семестре.

**Рисунок 2 –**  
**Визуализация**  
**запроса на выборку**  
**пар дисциплин с**  
**нарушением**  
**последовательност**  
**и изучения с**  
**выводом**  
**качестве**  
**узлов в**  
**label**  
**названия**  
**дисциплин**



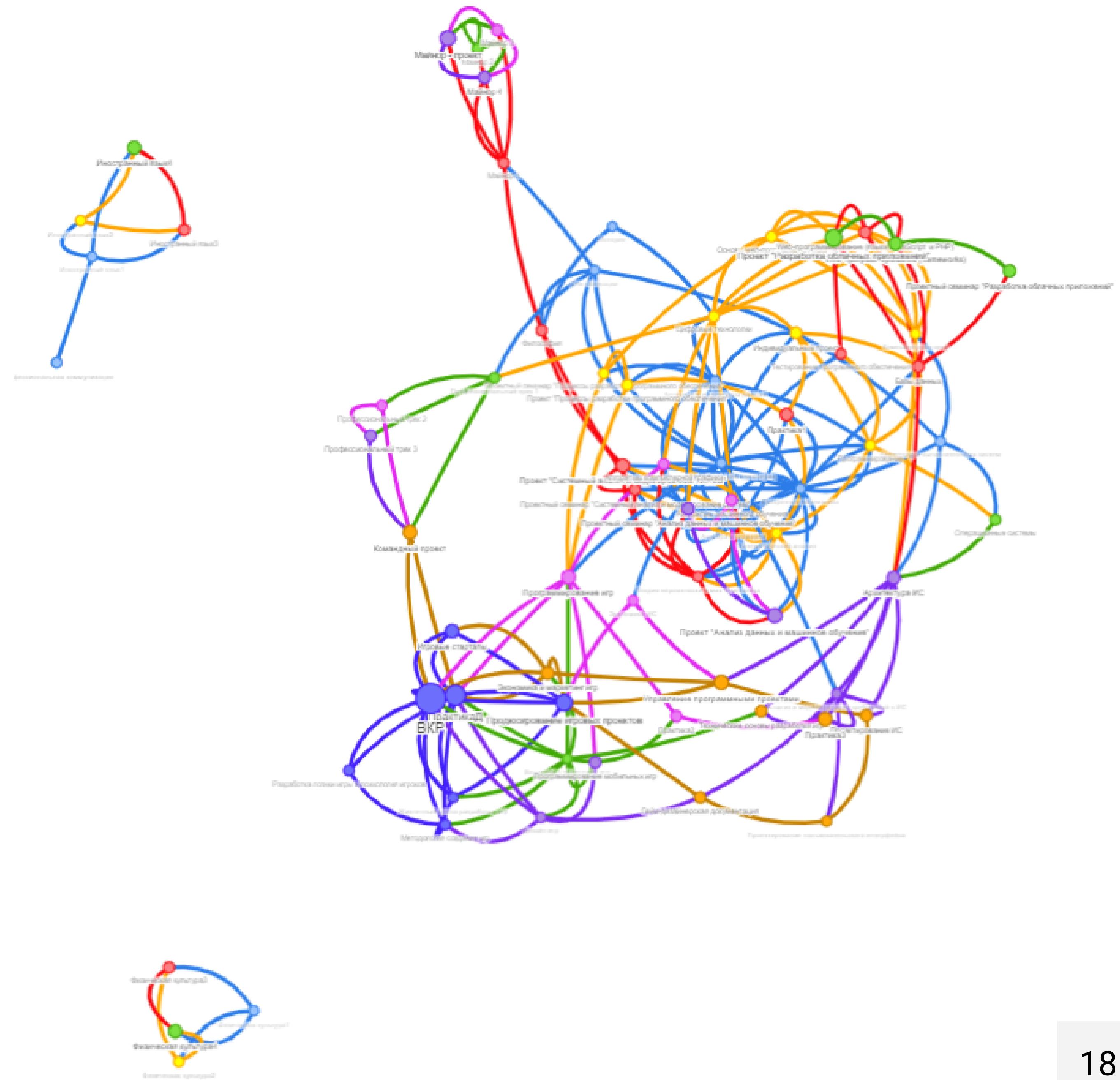
**Рисунок 3 –**  
**Визуализация**  
**запроса на выборку**  
**пар дисциплин с**  
**нарушением**  
**последовательност**  
**и изучения с**  
**выводом**  
**качестве**  
**узлов**  
**семестров**



# Результаты

В результате анализа логической последовательности изучения дисциплин было выяснено что среди всей выборки, 2966 пар дисциплин имеют нарушение последовательности изучения.

# Рисунок 4 – Визуализация запроса на выборку дисциплин учебного плана разрабатываемой образовательной программы. С визуализацией коэффициента PageRank диаметром узла, цветом узла и выходящих от него связей цветом соответствующим семестру изучения



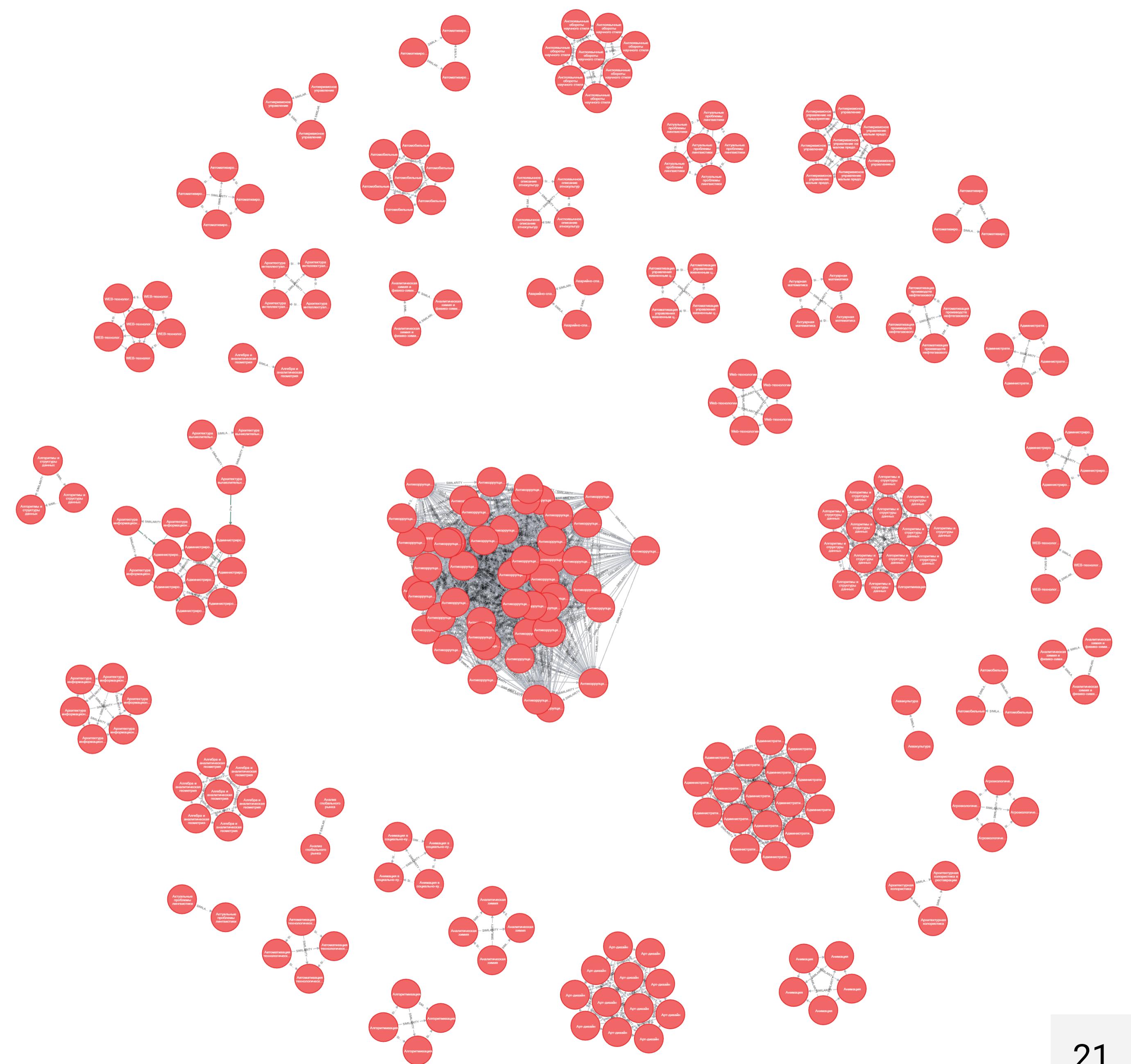
# Результаты

Как видно из рисунка 4, больший диаметр узла соответствует дисциплине «ВКР». Это означает что данная дисциплина имеет наибольший коэффициент транзитивного влияния (PageRank) среди всех дисциплин данной выборки.

# Кластерный анализ

Кластерный анализ проводился используя модифицированный алгоритм MinHash применяя попарное сравнение текстов РПД (поблочно, каждую с каждой)

# Рисунок 5 – Визуализация запроса на выборку дисциплин с имеющим связь типа *similarity* с порогом $> 90\%$ и лимитом на вывод в 300 улов



# Результаты

В ходе выполнения кластерного анализа было установлено значительное количество похожих по содержанию РПД при пороге схожести 90%. Среди отдельно взятых кластеров были обнаружены дисциплины с разными названиями. Порог кластеризации экспертами еще не установлен, в связи с этим нет точной информации о количестве кластеров. Как видно из рисунка 5 многие кластеры образуют полный граф, что косвенно говорит о правильности выбора порогового значения схожести для кластеризации.

# Итого

- Обнаружена **архитектурная ошибка** информационной системы для автоматизации управления учебным процессом лаборатории ММИС
- Согласованными по отношению к связям по типу пререквизит оказалось **28% связей**;
- Согласованными по отношению к связям по типу постреквизит оказалось **30% связей**;
- **2966 пар дисциплин** имеют нарушение последовательности изучения;
- Выявлены для каждого учебного плана **наиболее «важные» дисциплины**;
- Ожидаемое количество кластеров **на порядок меньше** количества РПД в базе данных и как минимум в несколько раз меньше количества РПД уникальных по названию.