НИЦ ИВТ

Ростех

Выявление аномалий в графах с использованием программной платформы Apache Spark

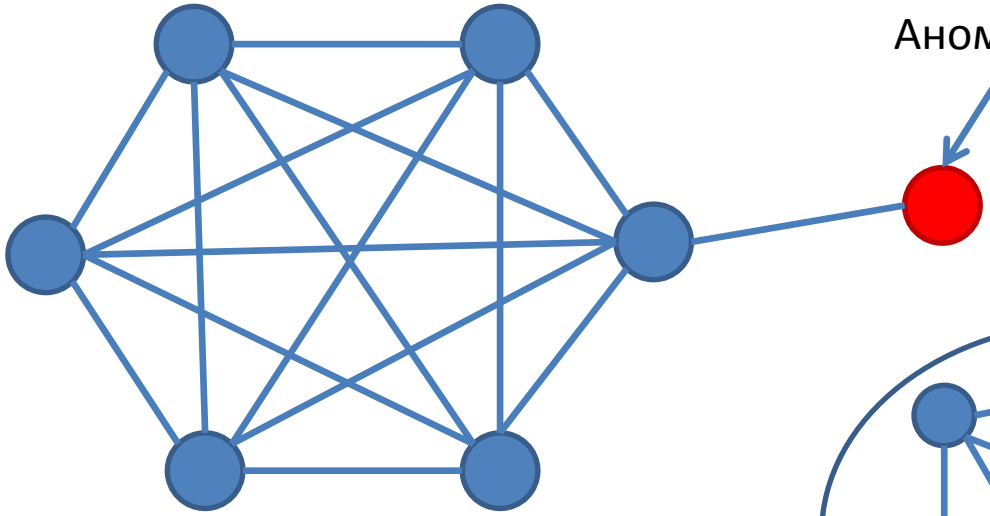
А.В. Мазеев, А.С. Семенов, Д.И. Дорофеев

1.03.2018

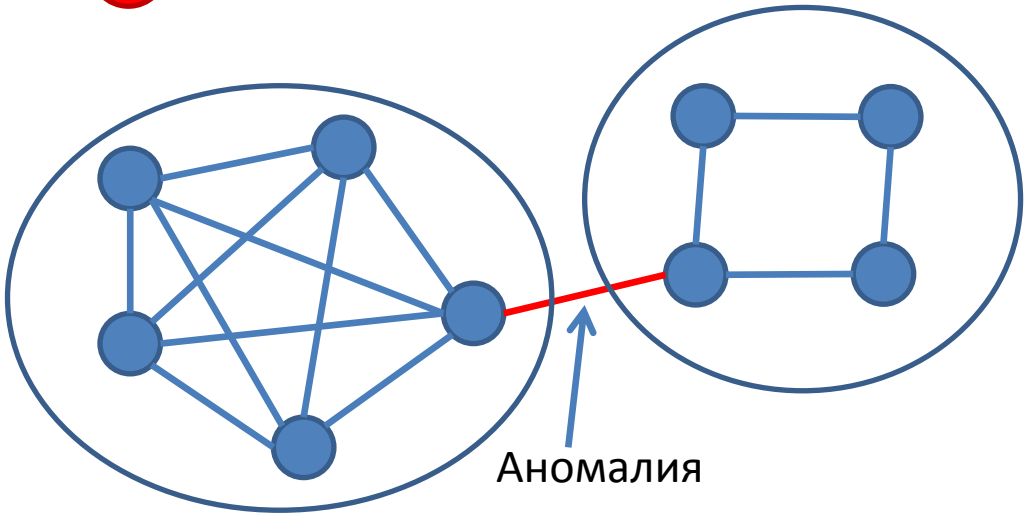
План

- Обзор методов выявления аномалий
- Выявление аномалий в реальном графе

Проблема



Аномалия



Аномалия

- **Дано:** граф G(V, E)
- **Задача:** найти аномалии в G – объекты (ребра, вершины, подграфы), которые существенно отличаются от большинства других объектов

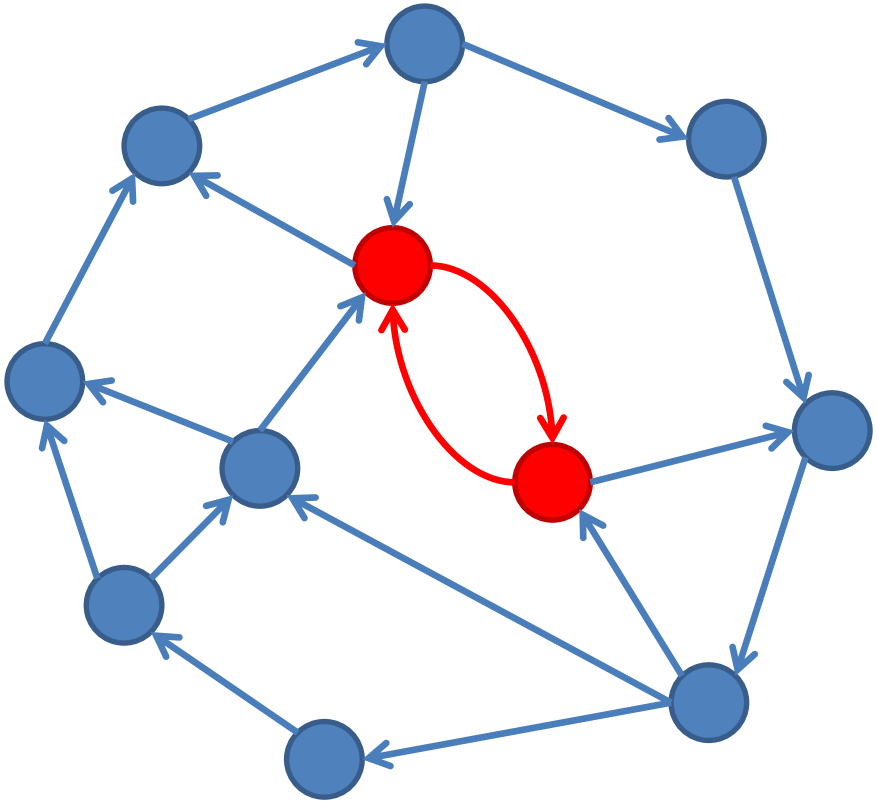
Актуальность

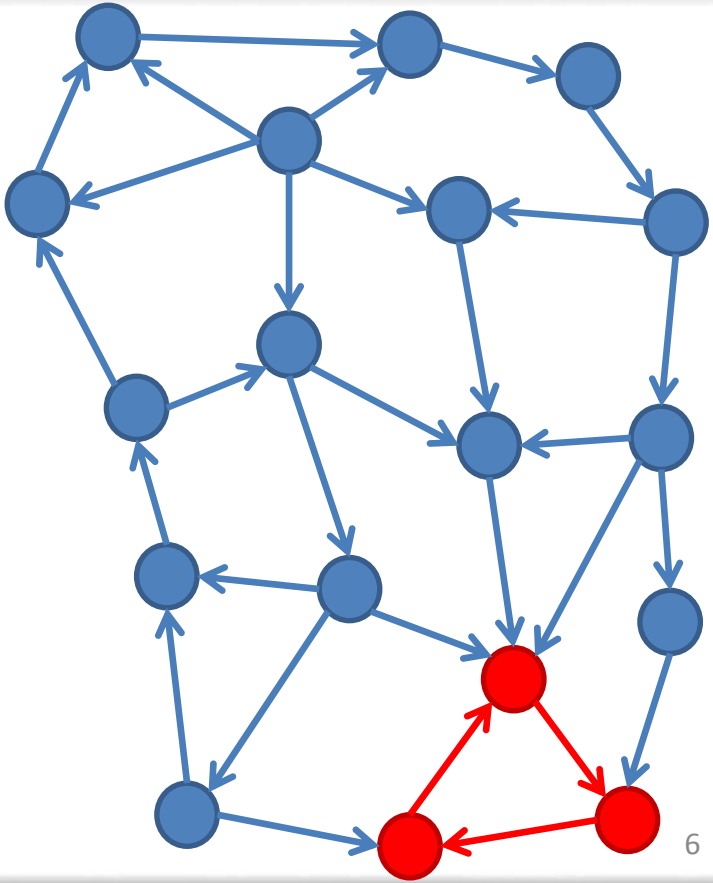
- Финансовые рынки
- Обнаружение спама
- Обнаружение кибератак
- Аномалии в социальных сетях, графах звонков

Рассматриваемые подходы

- На основе структур
- На основе сообществ (кластеризации)
- На основе признаков

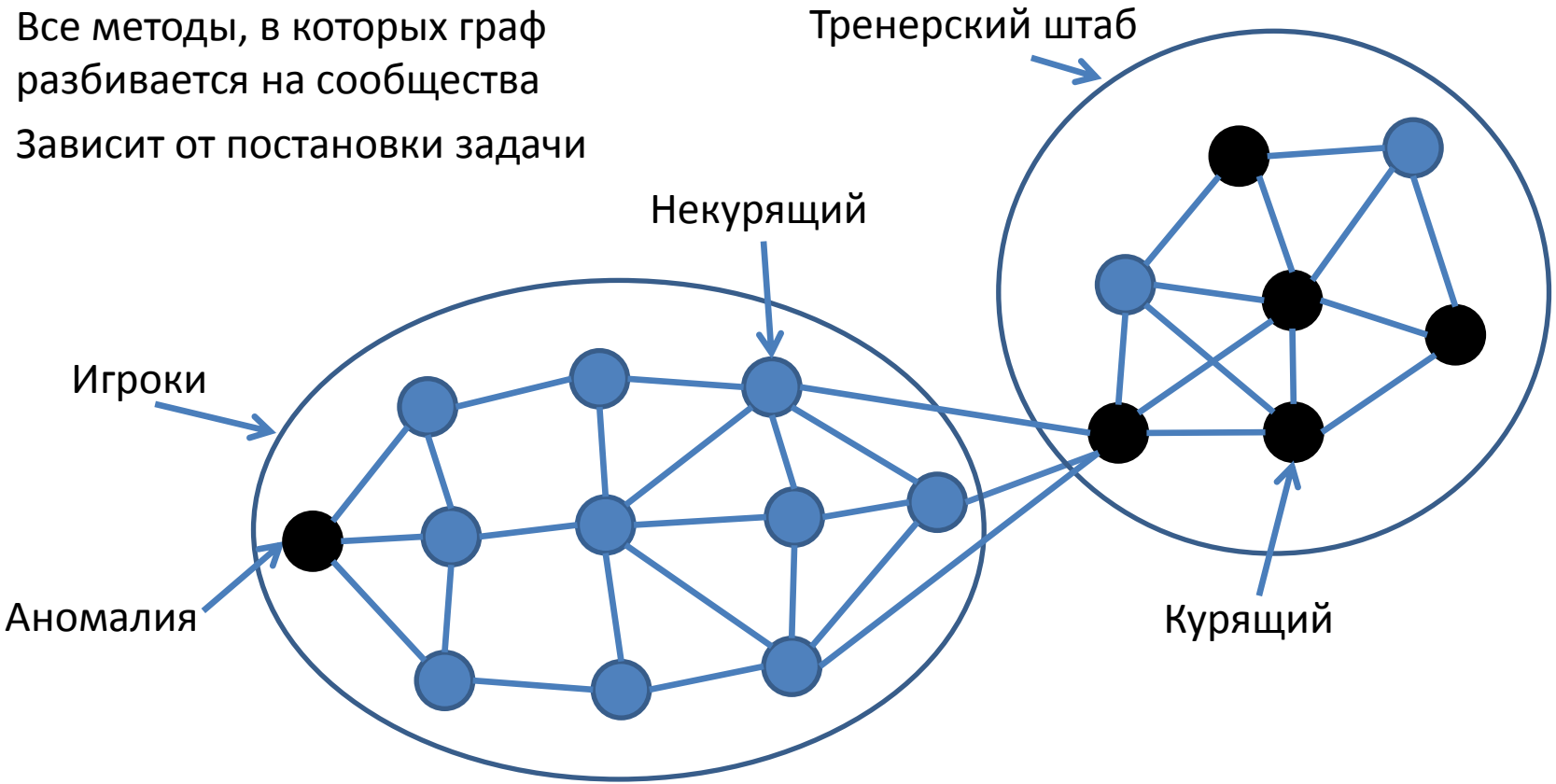
На основе структур





На основе сообществ

- Все методы, в которых граф разбивается на сообщества
- Зависит от постановки задачи



На основе признаков

- Для всех рассматриваемых объектов графа вычисляются некоторые числовые признаки (зависит от задачи)
- Тогда каждый рассматриваемый объект в графе можно представить в виде точки в n-мерном пространстве (вектор признаков)
- Далее среди этих точек необходимо найти выбросы (задача “outlier detection”)

На основе признаков
Признаки

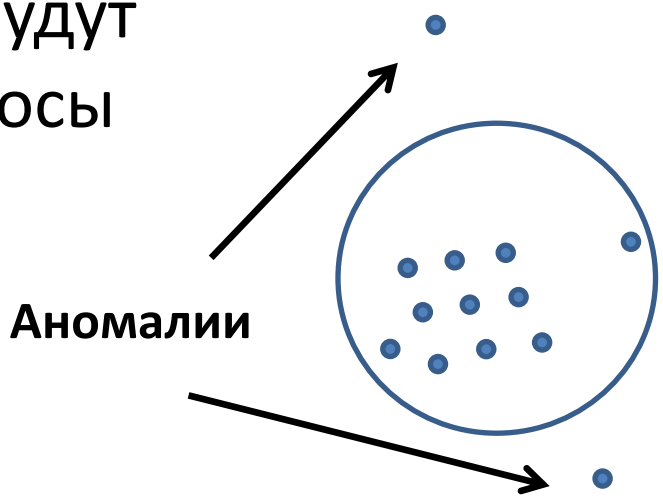
- Вершины (степень, полустепени, betweenness centrality, PageRank)
- Пары вершин (количество общих соседей, вес ребра)
- Egonet (количество треугольников, количество ребер)
- Группа вершин (плотность = кол-во ребер/кол-во вершин, общий вес ребер)

На основе признаков
Outlier detection

- Анализ распределения данных
- На основе расстояния до соседей
- На основе плотности

На основе признаков
Анализ распределения данных

- Найти центр масс
- Выбрать R, такое что точки на расстоянии больше R от центра масс будут рассматриваться как выбросы



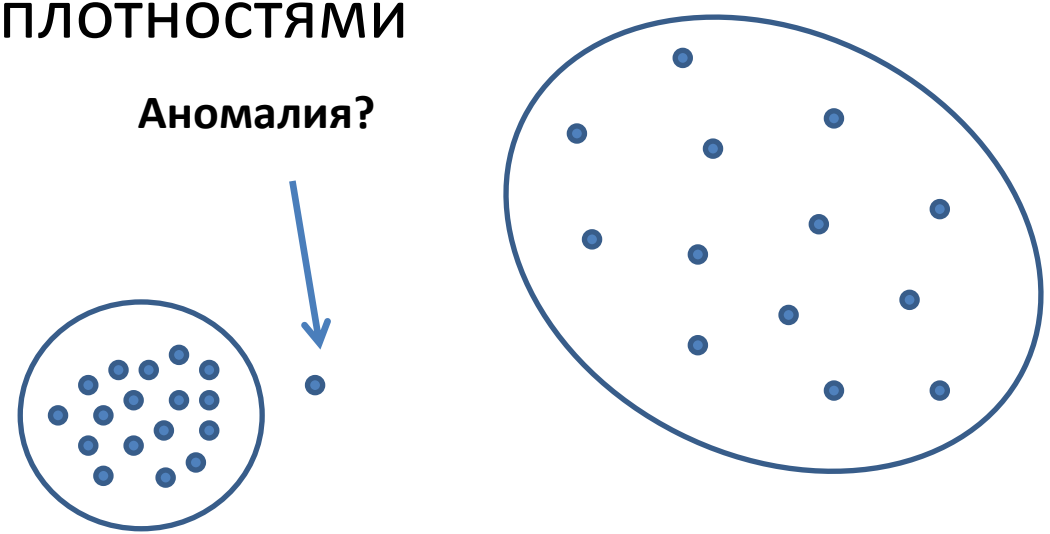
На основе признаков

На основе расстояния до соседей

- Два основных подхода:
  - Анализ среднего расстояния до k ближайших соседей (kNN)
  - Анализ количества соседей в радиусе ε (DB (ε, π))

На основе признаков
На основе плотности

- Сравнение плотности точки с плотностями соседей
- Метод на основе дистанций имеет проблемы с различными плотностями

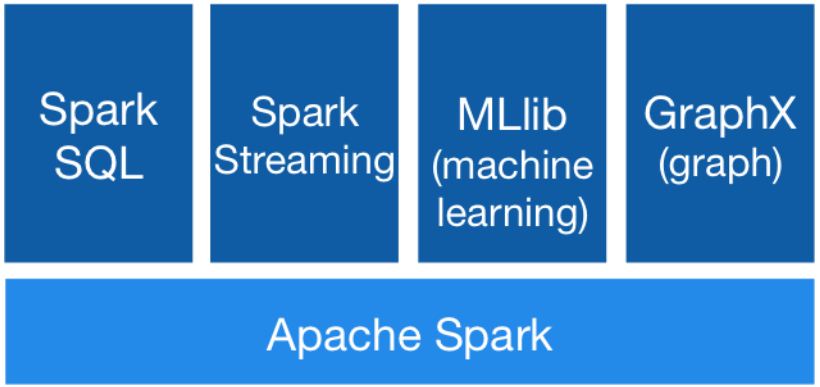


Описание задачи

- Дано: G(V, E, W) – взвешенный ориентированный граф
- Задача: найти аномальные ребра в G

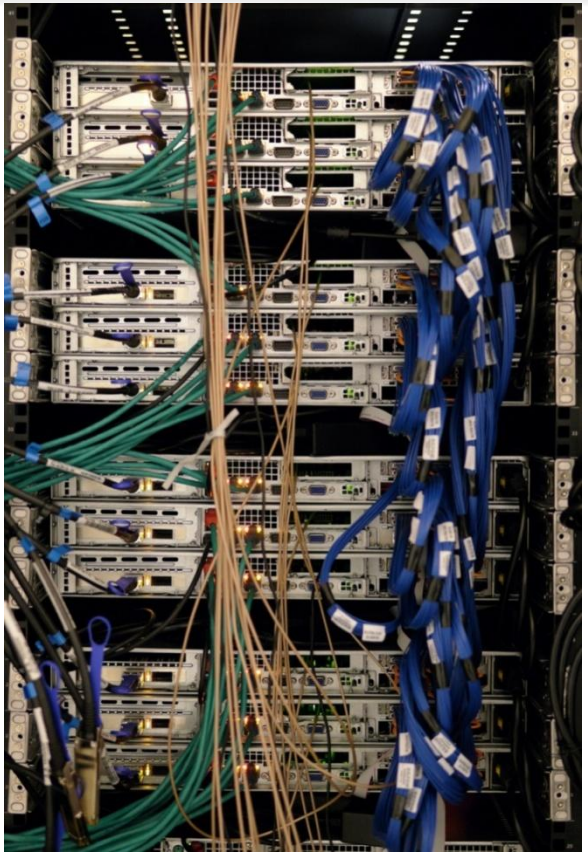
Scala + Spark

- Apache Spark — фреймворк с открытым исходным кодом для реализации распределённой обработки неструктурированных и слабоструктурированных данных



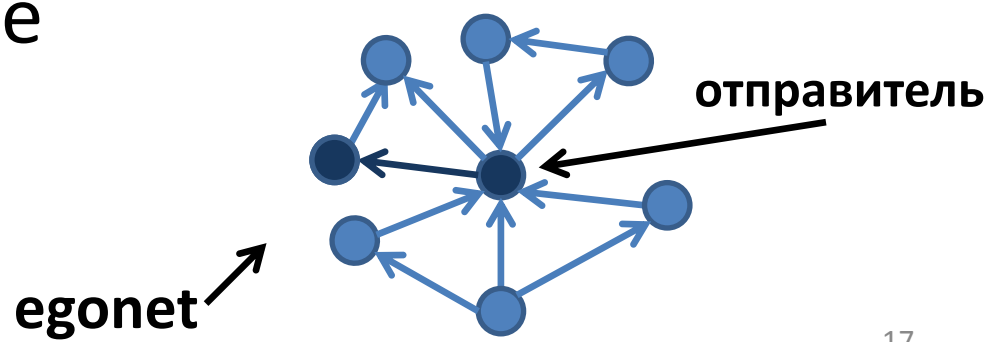
Кластер «Ангара-К1»

- **24 вычислительных узла**
  - Supermicro SuperServer 5017GR-TF
  - 2 процессора Intel Xeon E5-2630 (LGA2011, 6 ядер, 2.3 ГГц)
  - 64 ГБ
- **12 вычислительных узлов**
  - Supermicro SuperServer 5017GR-TF
  - процессор Intel Xeon E5-2660 (LGA2011, 8 ядер, 2.2 ГГц)
  - 64 ГБ
- **Сеть «Ангара»**
  - Адаптер EC8430, топология 3D-тор 3x3x4
  - Собственная реализация SHMEM
  - MPI: MPICH 3.0.4
- **Операционная система**
  - SLES 11 SP4, Linux 3.0.13-0.27-default



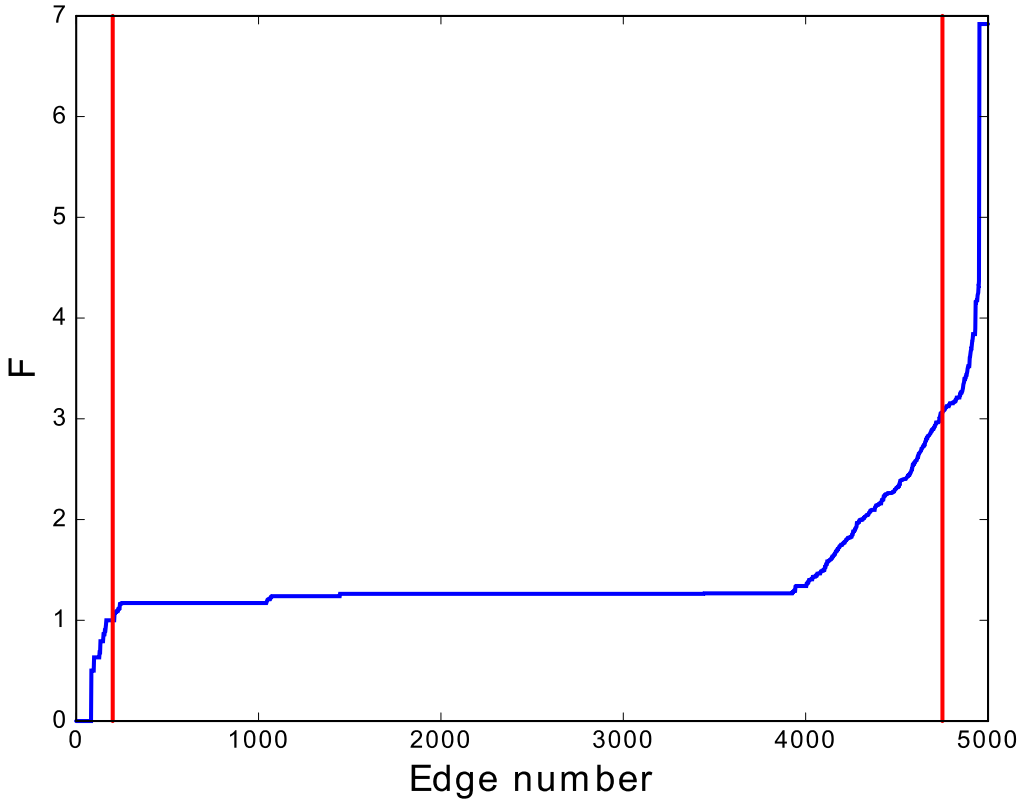
Метод на основе структур/признаков

- Для каждого ребра графа у вершины-отправителя выделяем egonet
- Пусть в i-ом egonet N_i вершин и E_i ребер
- Если $\alpha < F(N_i,E_i) < \beta$, то ребро нормальное, иначе – аномальное



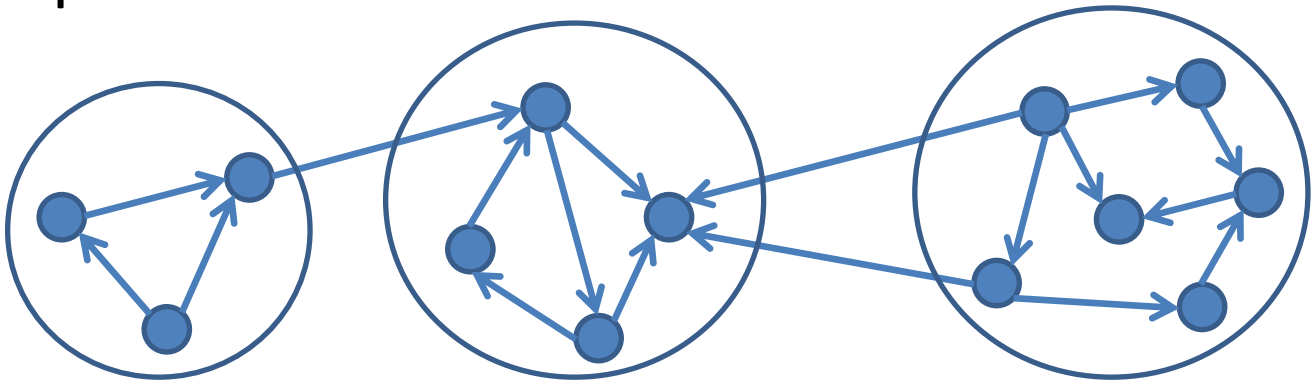
Метод на основе структур/признаков
Распределение

- $F(N_i,E_i) = \log_{N_i}(E_i)$
- Сортируем ребра по этой метрике
- Анализируем распределение
- Подбираем границы для выделения аномалий



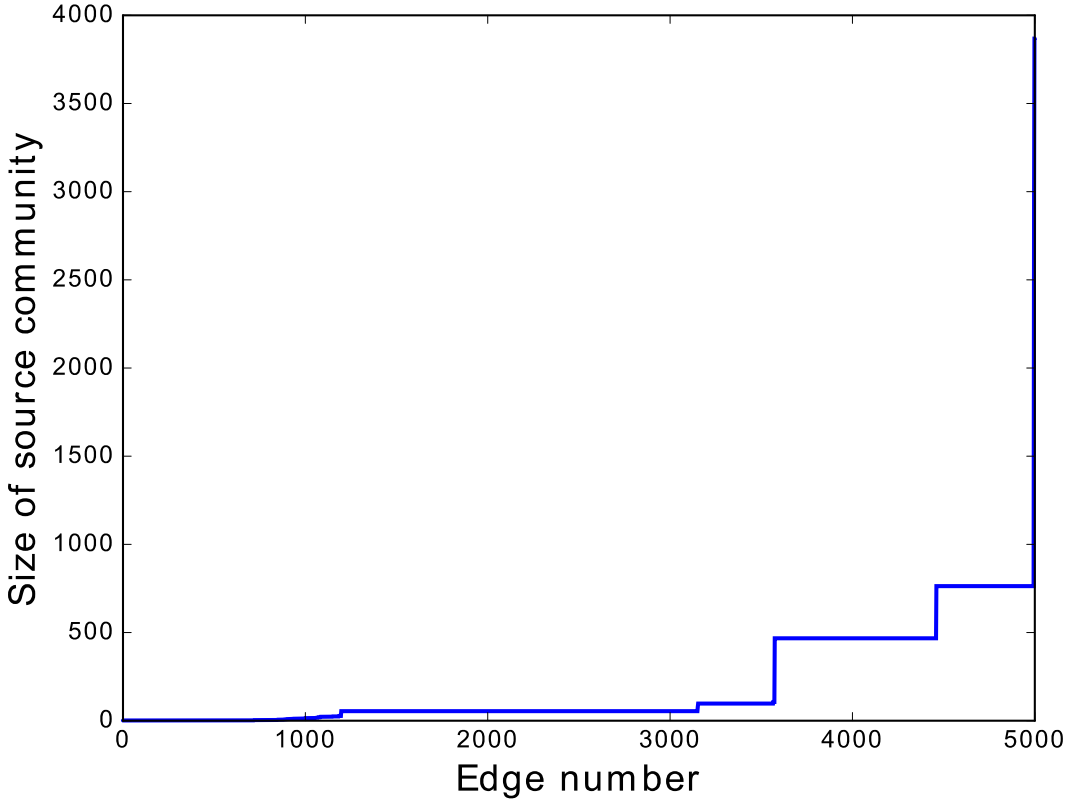
Метод на основе сообществ

- Разбиваем G на непересекающиеся сообщества (LabelPropagation)
- Для каждого ребра рассмотрим размер сообщества, в котором находится вершина-отправитель



Метод на основе сообществ
Распределение

- Сортируем ребра по размеру сообщества вершины-отправителя
- Анализируем распределение
- Подбираем границы для выделения аномалий



Примечание: те же самые ребра, что и в пред. методе

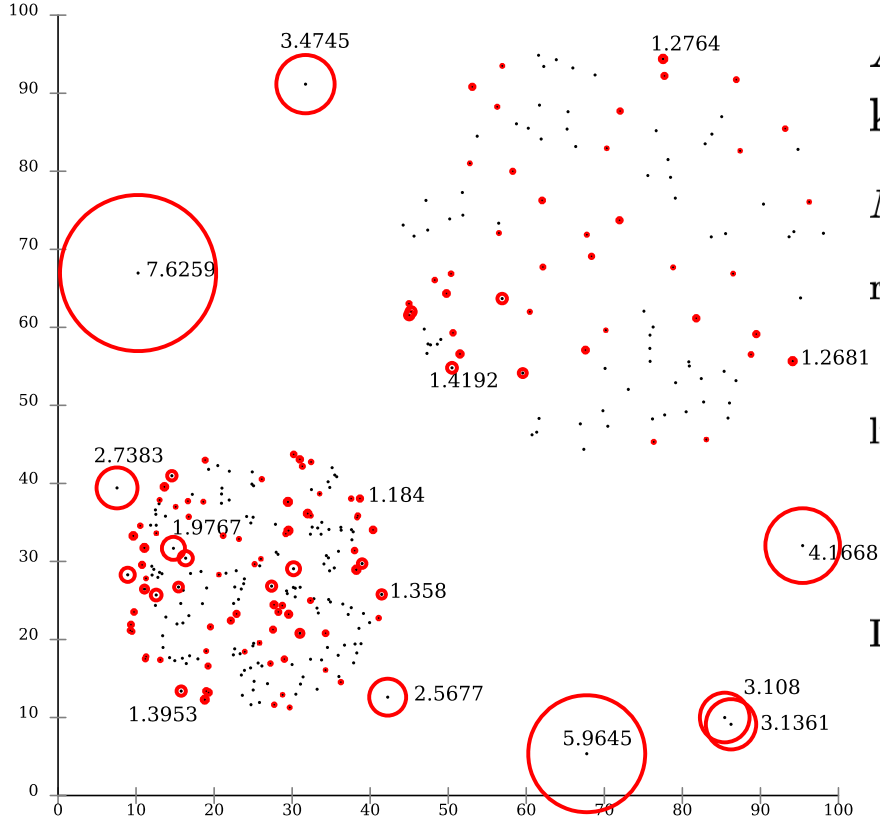
Метод на основе признаков

- Для каждого ребра вычисляем n числовых признаков
- Теперь ребра можно представить как точки в n-мерном пространстве
- Ищем выбросы с помощью метода LOF

Метод на основе признаков
Признаки (всего 37)

- Для отправителя и получателя:
  - Степень, полустепень исхода, захода
  - Для вершин в R=1
    - Кол-во вершин
    - Минимальная, максимальная и средняя степень, полустепень захода и полустепень исхода
    - Кол-во вершин-вулканов, вершин-черных дыр, оставшихся (обычных) вершин
  - Для ребер в egonet
    - Кол-во ребер
    - Сумма весов ребер
- Вес ребра

Метод на основе признаков
Local Outlier Factor, LOF



A - точка

$\text{k-distance}(A)$ - дистанция до k-го ближайшего соседа

$N_k(A)$ - множество k ближайших соседей

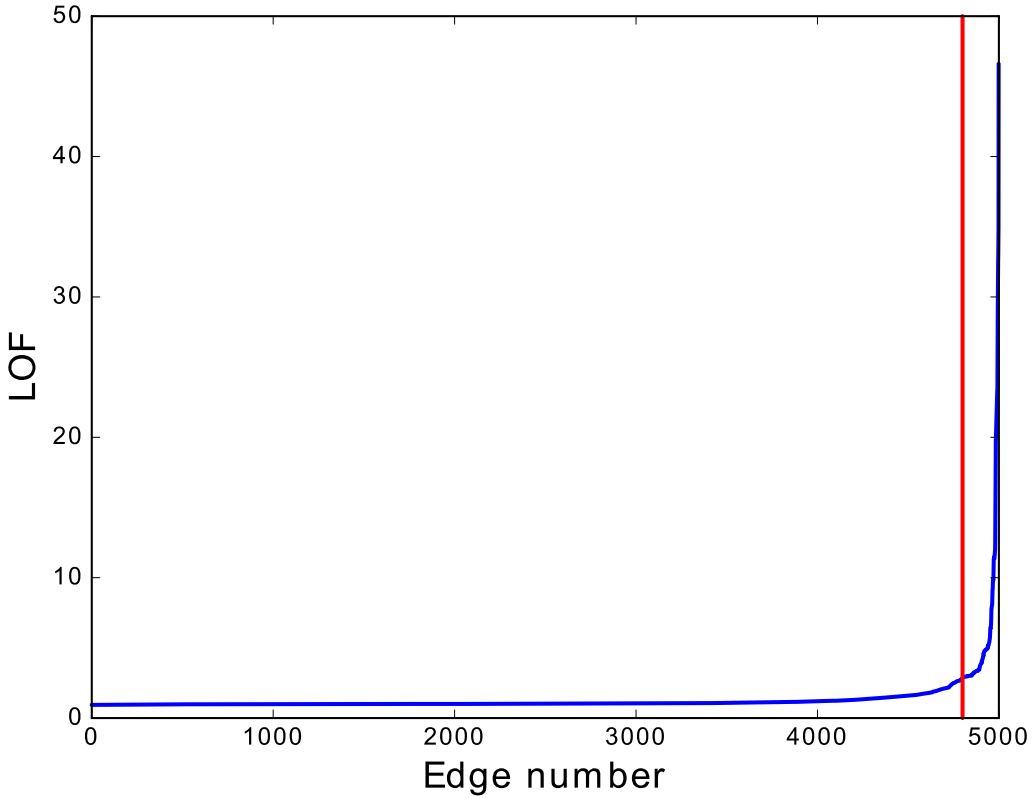
$$\text{reachability-distance}_k(A, B) = \max\{\text{k-distance}(B), d(A, B)\}$$

$$\text{lrd}(A) := 1/\left(\frac{\sum_{B \in N_k(A)} \text{reachability-distance}_k(A, B)}{|N_k(A)|}\right)$$

$$\text{LOF}_k(A) := \frac{\sum_{B \in N_k(A)} \frac{\text{lrd}(B)}{\text{lrd}(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} \text{lrd}(B)}{|N_k(A)|} / \text{lrd}(A)$$

Метод на основе признаков
Распределение

- Для каждого ребра получили значение LOF
- Сортируем ребра по этой метрике
- Анализируем распределение
- Подбираем границы для выделения аномалий



Примечание: те же самые ребра, что и в пред. методах

Результат

- Первый подход (плотность egonet): 450 аномалий (9%)
- Второй подход (LabelPropagation): 6 аномалий (0,12%)
- Третий подход (признаки+LOF): 200 аномалий (4%)
- Размер пересечения 1-го и 3-го подходов: 47 ребер

Заключение

- Были рассмотрены три метода выявления аномалий в графах
- Около 12% данных выявлены как аномальные
- Требуется их дальнейшая интерпретация

Настоящие исследования проводятся при финансовой поддержке Министерства образования и науки Российской Федерации. Уникальный ID прикладных научных исследований (проекта) RFMEFI57816X0218. Ответственность за представленные данные, результаты и выводы несут исключительно авторы исследования.

Спасибо за внимание!