

▼ Обработка данных

Чистила данные я прямо в Гефи, в Data Laboratory. Отсортировала Labels так, что стали видны сразу все Deleted, и я их удалила вручную. Там же я сделала красивые визуализации (лежат отдельными файлами в репозитории).

Дальше нужно было соединить таблицу узлов с таблицей ребер. Это получилось сделать в R.

```
library(tidyverse)
edges = read_csv('VK_Nika_edges.csv')
nodes = read_csv('VK_Nika_nodes.csv')
edges %>%
  left_join(nodes, by = c("Source" = "Id")) %>%
  left_join(nodes, by = c("Target" = "Id")) %>%
  rename(label_source = "Label.y", label_target = "Label") %>%
  select(Source, Target, Id, label_source, label_target) -> clean_graph
write.csv(clean_graph, 'clean_graph_data.csv') # здесь записываем данные в новый csv-файл
```

Переходим в Питон (Pycharm)

```
import networkx as nx
import collections
import matplotlib.pyplot as plt
import csv
import pandas as pd
from networkx import number_of_nodes
from networkx import number_of_edges
from networkx import density
from networkx import number_connected_components
from networkx import weakly_connected_components
from networkx import periphery
from networkx.algorithms.community centrality import girvan_newman
import community as louvain_community
import networkx.algorithms.community as nx_comm
from collections import Counter

data = pd.read_csv("clean_graph_data.csv")

labels = dict(zip(data['Source'], data['label_source']))

G=nx.from_pandas_edgelist(data, 'Source', 'Target')
pos=nx.spring_layout(G)
nx.draw_networkx_labels(G, pos, labels)
nx.draw(G, pos, with_labels=False, node_color='blue')
plt.show()
```



▼ Метрики

```
#Весь граф
print('Количество узлов:', number_of_nodes(G))
print('Количество ребер:', number_of_edges(G))
print('Плотность:', density(G))
print('Компоненты связности:', number_connected_components(G))
```

Количество узлов: 511
Количество ребер: 8048
Плотность: 0.0617627873067035
Компоненты связности: 1

```
print ('Диаметр: ', nx.diameter(G))
print ('Средняя длина пути: ', nx.average_shortest_path_length(G))
print ('Кластеризация: ', nx.transitivity(G))
print ('Кластерный коэффициент: ', nx.average_clustering(G))
print ('Количество узлов на периферии: ', len(nx.periphery(G)))
```

Диаметр: 6
Средняя длина пути: 2.53628026553087
Кластеризация: 0.3581921893646605
Кластерный коэффициент: 0.47380426736870446
Количество узлов на периферии: 34

График распределения степеней тоже автоматически выгрузила из Гефи, лежит отдельно в папке. Смысл его в том, что у самые большие значения degree - всего у нескольких человек, а у больших групп людей degree ниже. (Самое большое значение - 223 - у одного человека, который и является самым крупным узлом на графе, если ранжировать величину узлов по degree).

- ▼ Поиск сообществ разными алгоритмами

```
#Girvan-Newman algorithm
cm = nx_comm.girvan_newman(G)
communities = next(cm)
print(len(communities))
```

2

```
#Louvain algorithm
partition = louvain_community.best_partition(G)
modularity = louvain_community.modularity(partition, G)

print('Количество сообществ:', max(Counter(partition).values()))
```

➡ Количество сообществ: 6

Интерпретация

Гефи выдал очень хороший результат разделения на сообщества (оставила базовый resolution 1.0), все совершенно так, как и есть в жизни. Самый отдельный кластер - первая школа, где я училась до 9 класса. Видно, что вторая школа дала мне огромное количество новых связей, плавно перетекая в университет. Выделились даже более маленькие сообщества разных "тусовок".