[Алгоритмы на графах](https://otus.ru/learning/173921/?utm_source=email&utm_medium=email&utm_campaign=otus&utm_term=lesson_remind_new&relogin=True&token=1af75954f593f32502e519c063b078a583ce923f)

Домашнее задание

Набор данных: <https://www.kaggle.com/datasets/flashgordon/usa-airport-dataset?resource=download>

1. Препроцессинг

Данные помесячные. Выделены данные за 12 месяцев 2008 года. Всего в датафрейме 232109 записей. Определены списки городов отправления (415 городов), прибытия (418) и всех городов (427 - с некоторыми городами не было двустороннего авиасообщения).

1. Пишем симуляцию.

*Напишите функцию для симуляции одного прохода по всему датасету и распространения инфекции. Функция должна принимать на вход стартовый город, в котором изначально началась инфекция и вероятность распространения инфекции p. На выходе функция должна возвращать словарь, в котором ключом будет время заражения аэропорта, значением - название аэропорта. В теле функции вам нужно написать проход по всему датасету (имеет смысл использовать itertuples), где вы будете проверять, является ли текущий рейс “заразным” и если да - с вероятностью p заражать новый аэропорт. В качестве стартовой ноды/аэропорта для симуляций можно использовать Allentown (node\_id = 0).*

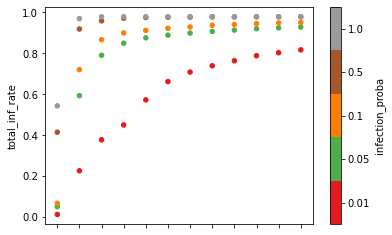
В качестве стартового взят **Aberdeen, SD** (первый в списке город отправления). Перебор датафрейма через itertuples работает довольно медленно, один проход датасета примерно за 30 сек.

Вместо словаря в результате одного прохода выдается DataFrame **inf\_hist\_** с данными о вероятности заражения и номере симуляции, который конкатенируется далее в единый датафрейм с результатами – **infection\_history**

Также, сразу считается статистика **inf\_stat** – число зараженных городов на временном шаге и их доля от числа всех городов.

Результаты сохранены в папку results, файлы MS Excel.

1. Как инфекционность (вероятность заражения) болезни влияет на скорость распространения?
   1. *Давайте посмотрим, насколько заразной должна быть болезнь, чтобы заразить все аэропорты в сети. Для каждой из вероятностей p=[0.01, 0.05, 0.1, 0.5, 1.] прогоните по 10 симуляций распространения, каждый раз сохраняя результаты.*
   2. Для каждых 12 часов симуляции посчитайте, какой средний процент аэропортов от общего их числа, был заражен к этому моменту времени. Постройте графики числа зараженных городов в зависимости от времени для каждой из вероятностей



При вероятности 0.5, 1 заражение идет очень быстро, в первый месяц заражено более 90% городов. Вероятность 0.01 и 0.05 дает более плавную картину заражения.

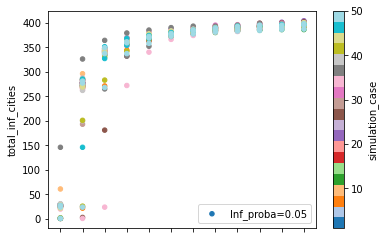
1. Может ли информация о сети помочь предсказать, кто будет заражен?
   1. *Настало время поработать с NetworkX. По имеющемуся датасету вам нужно построить ненаправленный граф всех аэропортов, используя относительное число перелетов между соседними аэропортами в качестве веса ребра. Например, если из аэропорта А в аэропорт Б суммарно есть 5 рейсов, из Б в А 10 рейсов, а всего в сети 150 перелетов, то вес ребра между А и Б должен быть равен (5+10)/150 = 0.1.*

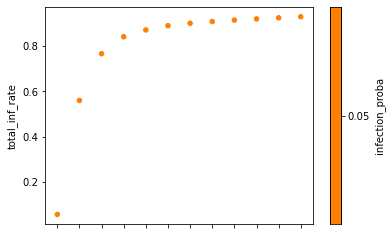
Перелеты между городами на каждый шаг посчитаны с помощью pd.groupby.

Веса ребер посчитаны как для каждой даты – **flights\_weighted\_by\_dates,** так и для всех дат - **flights\_all.**

* 1. *Прогоните вашу симуляцию 50 раз, используя p=0.5 и стартовый аэропорт Allentown, каждый раз сохраняя результаты. Используя данные симуляций посчитайте медианное время до заражения каждого города.*

В качестве стартового взят **Aberdeen, SD** (первый в списке город отправления). Вероятность взята равной 0.05, так как при 0.5 заражение 90% городов уже в первый месяц. Выполнено 50 симуляций.





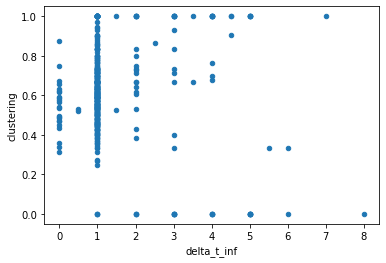
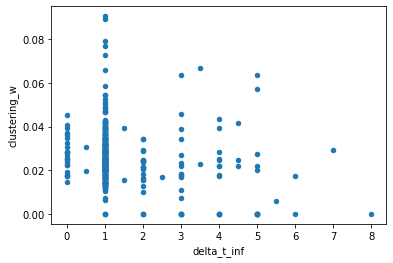
* 1. *Теперь используя построенный вами граф, для каждого города посчитайте его коэффициент кластеризации (nx.clustering), степень (nx.degree) и центральность (nx.betweenness\_centrality).*

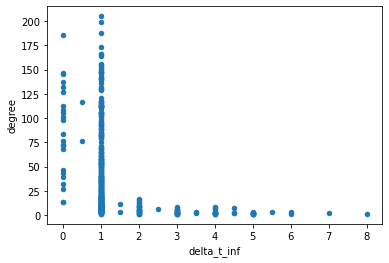
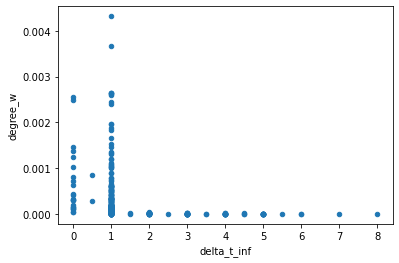
Все три метрики посчитаны как с весами, так и без весов. Все метрики для городов-узлов сведены в один датафрейм **inf\_graph.**

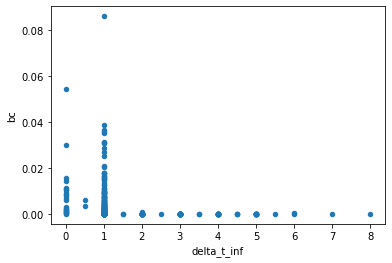
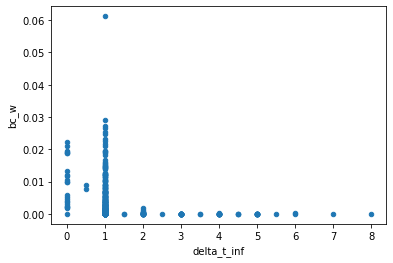
* 1. *Наконец, постройте scatter-plot-ы для медианного времени заражения и каждой из рассчитанных выше метрик. Также посчитайте коэффициент корреляции Спирмана между этими показателями и попробуйте проинтерпретировать полученные результаты. Какая из метрик графа сильнее всего скоррелирована со временем заражения? Почему?*

На графиках: Число месяцев от момента заражения – **delta\_t\_inf.**

Метрики графа (c \_w – взвешенные по общему числу перелетов) – **clustering, clustering\_w, degree, degree\_w, bc, bc\_w**

Корреляция Спирмана:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Столбец1** | **delta\_t\_inf** | **clustering** | **degree** | **bc** | **clustering\_w** | **degree\_w** | **bc\_w** |
| **delta\_t\_inf** | 1.00 | 0.13 | -0.61 | -0.58 | -0.22 | -0.63 | -0.56 |
| **clustering** | 0.13 | 1.00 | -0.29 | -0.47 | 0.57 | -0.28 | -0.38 |
| **degree** | -0.61 | -0.29 | 1.00 | 0.94 | 0.19 | 0.99 | 0.91 |
| **bc** | -0.58 | -0.47 | 0.94 | 1.00 | 0.05 | 0.93 | 0.95 |
| **clustering\_w** | -0.22 | 0.57 | 0.19 | 0.05 | 1.00 | 0.26 | 0.04 |
| **degree\_w** | -0.63 | -0.28 | 0.99 | 0.93 | 0.26 | 1.00 | 0.89 |
| **bc\_w** | -0.56 | -0.38 | 0.91 | 0.95 | 0.04 | 0.89 | 1.00 |

Большие значения, выделенные красным – 0.89 – очевидно связаны с нулевыми значениями на шагах 2-12.

Зеленым выделены коэффициенты, которые могли бы свидетельствовать об обратной зависимости показателей.

* Степень посредничества bc должна быть выше у городов - главных узлов сети аэропортов, с максимальным траффиком, через которые ведут пути ко всем другим аэропортам и следовательно более быстрым заражением
* Степень вершин **degree** также должна быть связана с временем заражения обратной зависимостью (меньше связей – дольше время до заражения)
* Кластеризация **clustering** – заражение в кластере должно происходить в одно время, низкая кластеризация – позднее время заражения.