Исследование графов взаимодействий павианов

Морозов Евгений, Сурков Антон, Ведерникова Мария

Что было сделано

1) Отображение и анализ графов взаимодействий павианов в статике и в динамике, выделение особенностей конкретных особей и связей между ними [ноутбук **Визуальный анализ**]

2) С помощью рандомного выбора подграфа для графов с разными типами поведения (условно дружественное и враждебное) показаны различия различия коэффициентов кластеризации, среднего числа компонент связности графов для разных категорий поведения. Также добавлено сравнение со случайными графами с аналогичными параметрами. Поставлена, но не принята гипотеза о связи степени сходства графов взаимодействий за разные дни с тем, насколько эти дни близки друг к другу [ноутбук **Сравнение графов**]

3) Сравнение различных способов выделения важности вершин: на основе анализа частот и времени взаимодействий, а также на основе различных centrality алгоритмов [ноутбук **Community detection**]

4) С помощью данных о контактном взаимодействии смоделированы различных типов эпидемий (SI, SIS, SIR) с разными параметрами и проведен анализ полученных результатов [ноутбук **Эпидемии**]

Основные результаты будут приведены ниже, подробные реализации и нюансы можно смотреть в ноутбуках. **Особенно многочисленные изображения графов!** Множество функций, связанных с моделированием, обработкой и отображением данных, вычислением параметров и прочим, вынесены в отдельные модуль **functions.py**.

Описание данных

Источник данных - датасет Baboons interactions, содержащий в себе а) данные наблюдений за группой из 20 Гвинейских павианов б) данные носимых павианами сенсоров. Обе группы данных относятся к одному периоду: от 13 июня до 10 июля 2019 года.

Файл наблюдений (OBS_data.txt) содержит описание элементов поведения, зарегистрированных наблюдателем. Он содержит 7 столбцов:

- Datatime - дата и время события
- Actor - имя особи, проявившей ту или иную форму поведения
- Recipient - имя особи, в отношении которой было проявлено поведение
- Behavior - типы проявленного поведения Выделялись 14 типов: ‘Presenting’,’Playing with’, ‘Grunting-Lipsmacking’, ‘Supplanting’,’Threatening’, ‘Submission’, ‘Touching’, ‘Avoiding’, ‘Attacking’,’Carrying’, ‘Embracing’, ‘Mounting’, ‘Copulating’, ‘Chasing’, ‘Invisible’ and ‘Other’
- Category - признак, квалифицирующий тип поведения. Выделяли три категории: ‘Affiliative’, ‘Agonistic’, ‘Other’;
- Duration - продолжительность проявления поведения
- Point - индикатор того, относится ли событие к категории POINT event (указывается значение “YES”) или STATE event (указывается значение “NO”).

Файл содержит 3196 парных наблюдений.

В первом датасете мы концентрировались именно на парных взаимодействиях, то есть тех, где указан Recipient. Для удобства отображения графов имена павианов заменены на номера.

Файл данных сенсоров (RFID_data.txt) содержит данные по 13 из 20 павианам за тот же период в 4 столбцах:

- t - время начала контакта в формате Epoch
- i - имя первой особи
- j - имя второй особи
- Datetime - дата и время в формате datetime

Файл содержит 63095 событий.

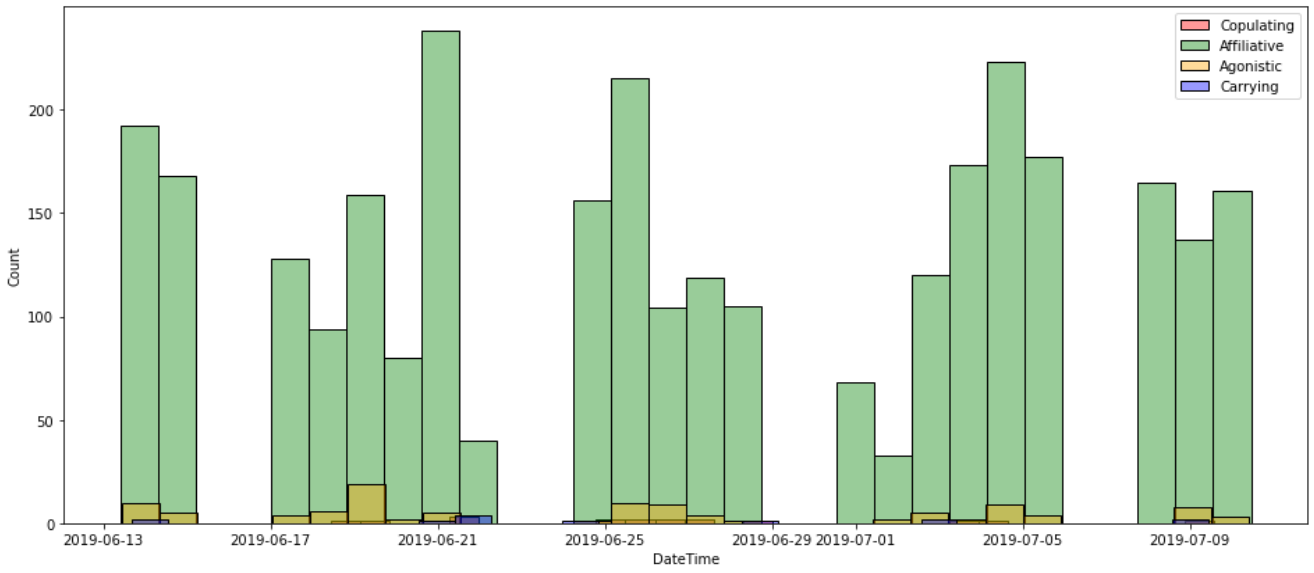
Визуальный анализ

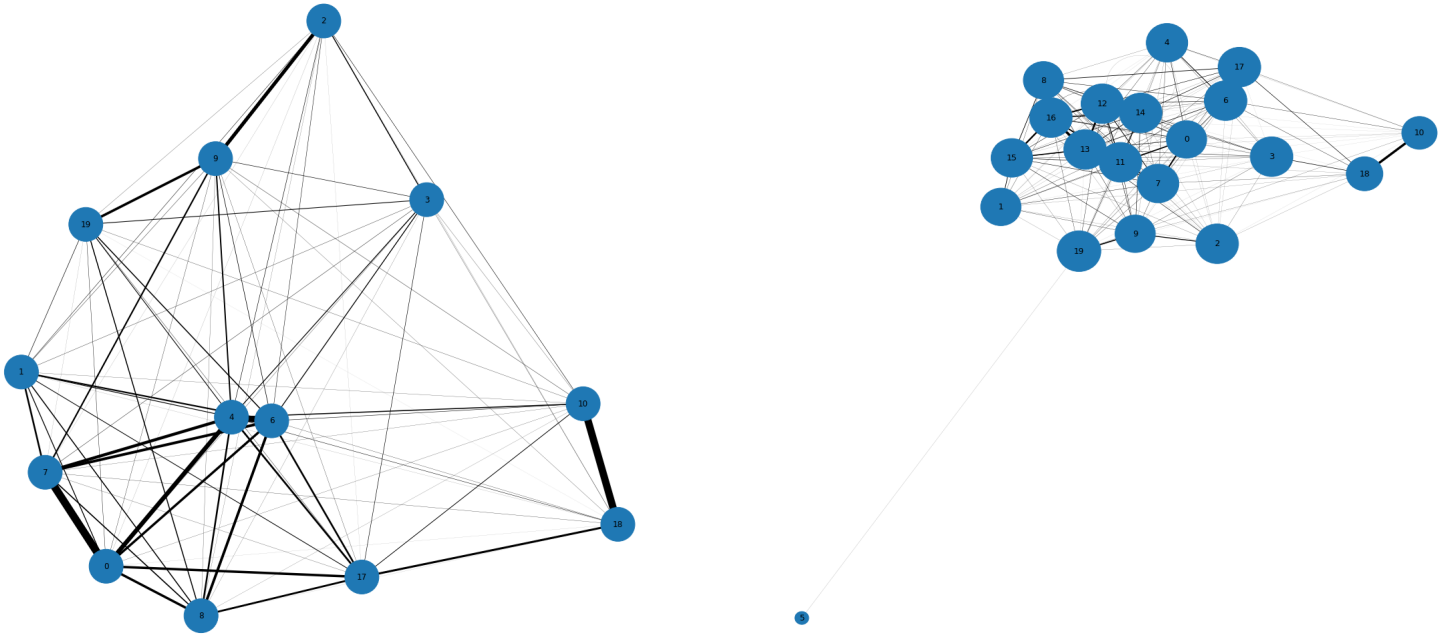
Файл OBS содержит следующее распределение видов поведения:

Category	Behavior	
Affiliative	Embracing	58
	Grooming	438
	Grunting-Lipsmacking	142
	Mounting	30
	Playing with	771
	Presenting	215
	Resting	1247
	Touching	156
Agonistic	Attacking	10
	Avoiding	5
	Chasing	19
	Submission	36
	Supplanting	14
	Threatening	25
Other	Carrying	13
	Copulating	17

Как можно заметить, аффилиативного взаимодействия намного больше, чем агонистического. Существенную его часть составляют отдых (по описанию в оригинальной статье, для парных взаимодействий речь идет именно о социальном отдыхе, когда обезьяны делают это вместе; отдых в одиночестве в изначальной таблице также присутствует, но здесь не отображен), груминг и игра.

Агонистических взаимодействий меньше. В отдельную категорию выделены совокупление и ношение. Вот так наблюдения в различных категориях распределены по дням:



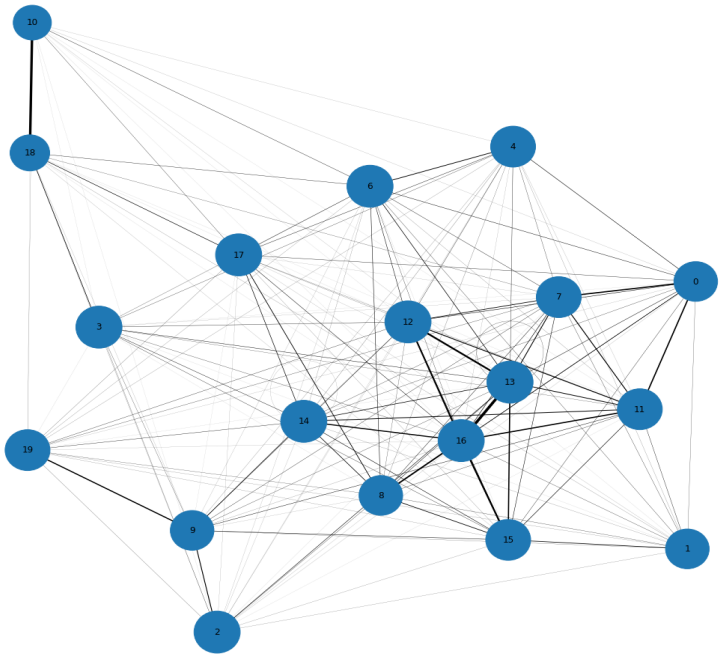


Общая структура взаимодействий в графах RFID (слева) и OBS (справа) за все время

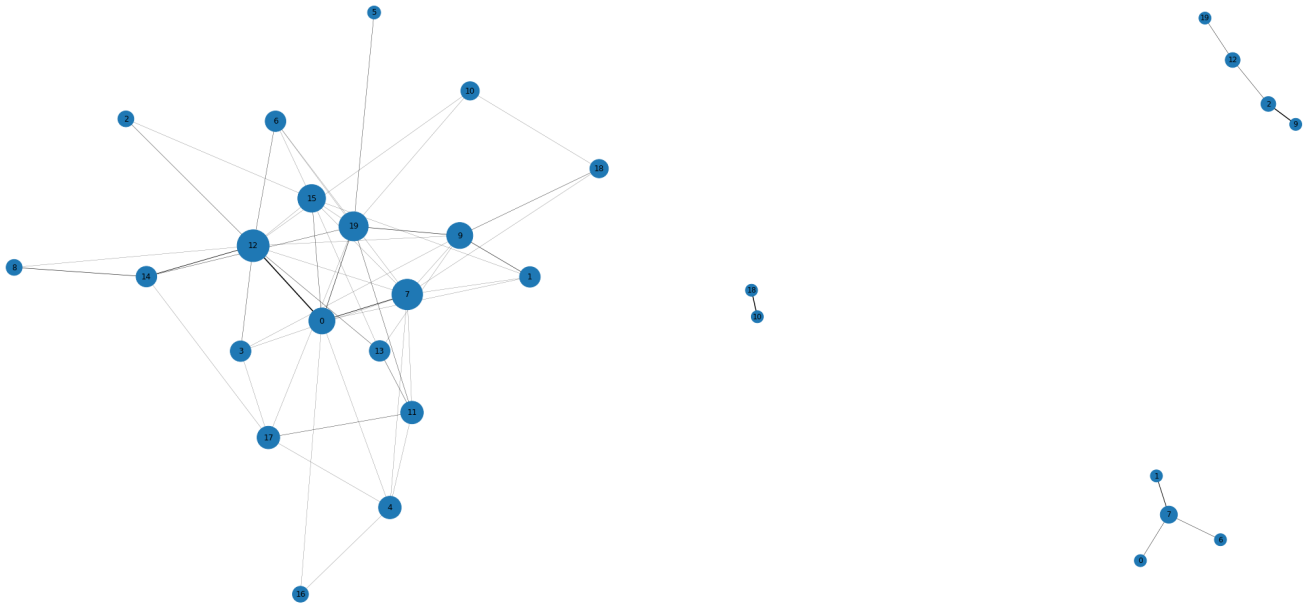
В RFID представлены не все особи - сенсоры носили только 13 из 20. Из этих картинок можно сделать следующие выводы:

1) Павиан №5 почти не участвует во взаимодействиях. Его имя EXTERNE, должно быть, выбрано не случайно

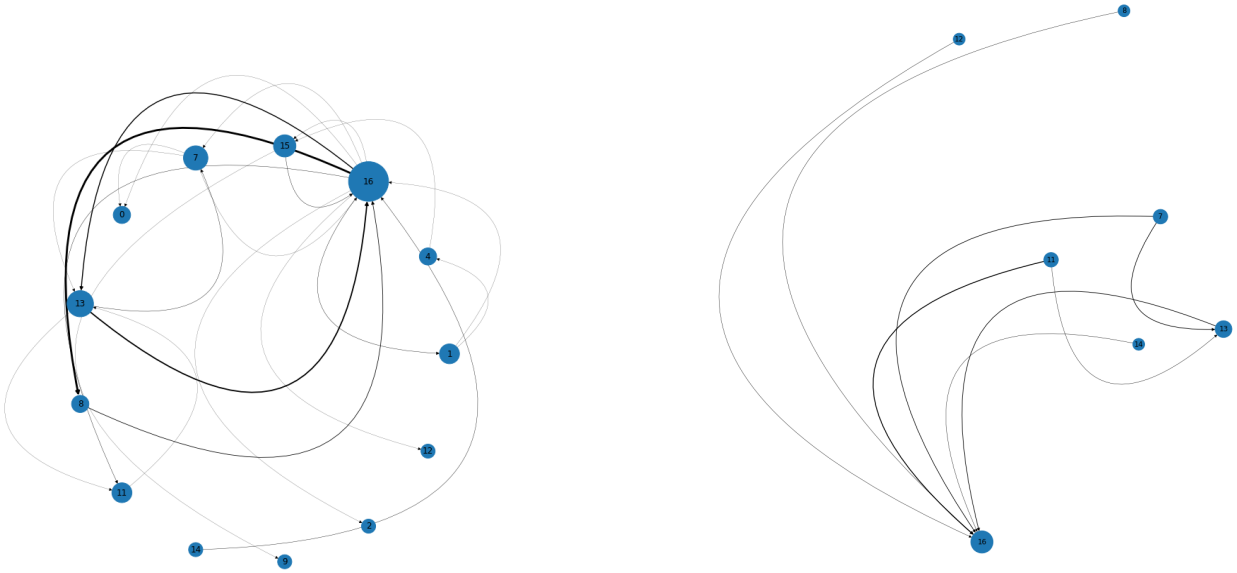
2) В OBS есть сильно связанные между собой и слабо с другими особи 10 и 18. Также есть плотно связанная между собой группа 11-16. В RFID значимость взаимодействий 10-18 тоже высока, но появляются и другие ребра с большим весом (вес пропорционален частоте взаимодействий) - 2-9, 4-6, 4-0, 0-7. То есть сенсоры демонстрируют плотный контакт в то время как наблюдения - нет. Возможно, это связано с тем, что не все особи в любой момент находятся под наблюдением.



Слева показаны аффилиативные взаимодействия (граф похож на OBS в целом)



Агонистические взаимодействия (центральные вершины - 0, 7, 12, 19) слева, граф совокуплений справа (тут не видно, но цепочка вершин 19-12-2-9, отдельная пара 10-18 и в оставшемся трезубце в центре 7, а по краям 0, 1, 6).



Граф объятий слева, граф ношения справа. И там, и там самые большие входящие степени у 13 и 16 вершин. Все носят 13 и 16 и только их! Обнимают тоже преимущественно именно их. По совокупности наблюдений (как много аффилиативного взаимодействия, в частности объятий и ношения, они получают, как мало агонистического, а также их неучастия в совокуплении) можно выдвинуть гипозету, что 13 и 16 - дети.

Также можно отметить пару 18-10, которая представлена и на графе совокуплений, и тот факт, что активные вершины графа совокуплений и активные вершины агонистического графа очень схожи между собой.

Сравнение графов

Идея данного раздела такая: хочется найти различия в структуре графов аффилиативного и агонистического взаимодействия, а также сравнить их с рандомными. Эти графы (мультиграфы, на самом деле) имеют разную реберную мощность, поэтому параметры у них отличаются (например, в аффилиативном ребер так много, что его коэффициент кластеризации около 0.9).

Моделирование проводилось следующим образом: фиксировалось количество вершин (брались все 20) и количество ребер (20, 35, 50). Ребра случайным образом выбирались из соответствующих мультиграфов, а для случайного - из всего возможного множества ребер, с повторениями.

Эксперимент повторялся 100 раз, каждый раз вычислялись разные функции и сравнивались полученные наборы.

К сожалению, большинство мер, к примеру, betweenness, так не посчитать, потому что они требуют связности. Так что были вычислены коэффициент кластеризации и среднее число компонент связности для всех вариантов количество ребер и для всех вариантов графов (агонистический, случайный, аффилиативный).

	agonistic	random	affiliative	p-value
average_clustering 20	0.05398	0.05181	0.09018	8.398742e-06
average_clustering 35	0.12780	0.13523	0.15789	4.316138e-03
average_clustering 50	0.21319	0.22532	0.25089	6.440531e-04
number_connected_components 20	6.11000	4.25000	6.15000	1.296676e-23
number_connected_components 35	3.16000	1.58000	3.43000	5.895633e-34
number_connected_components 50	1.91000	1.11000	2.56000	2.542166e-35

Результаты показаны в таблице выше. В самом правом столбце p-value - результат теста one-way ANOVA для всех трех выборок. Впрочем, возможно, его стоило сделать и попарно.

По таблице можно видеть, что средний коэффициент кластеризации у агонистического графа ниже случайного, а у аффилиативного выше. По смыслу это выглядит очень логично - дружественное взаимодействие легко предполагает треугольники, враждебное - скорее нет.

Различия в количестве компонент связности таковы: у агонистических графов их чуть поменьше, чем у аффилиативных, но у обоих намного больше, чем в случае рандома. Это означает, что агонистические и аффилиативные взаимодействия происходят (как ни странно) не случайно и потому чаще повторяются в определенных группах (мы видели эти тяжелые ребра в визуальном анализе). Чуть большая связность агонистических графов по сравнению с аффилиативными же объясняется, по-видимому, той же самой повышенной кластеризацией

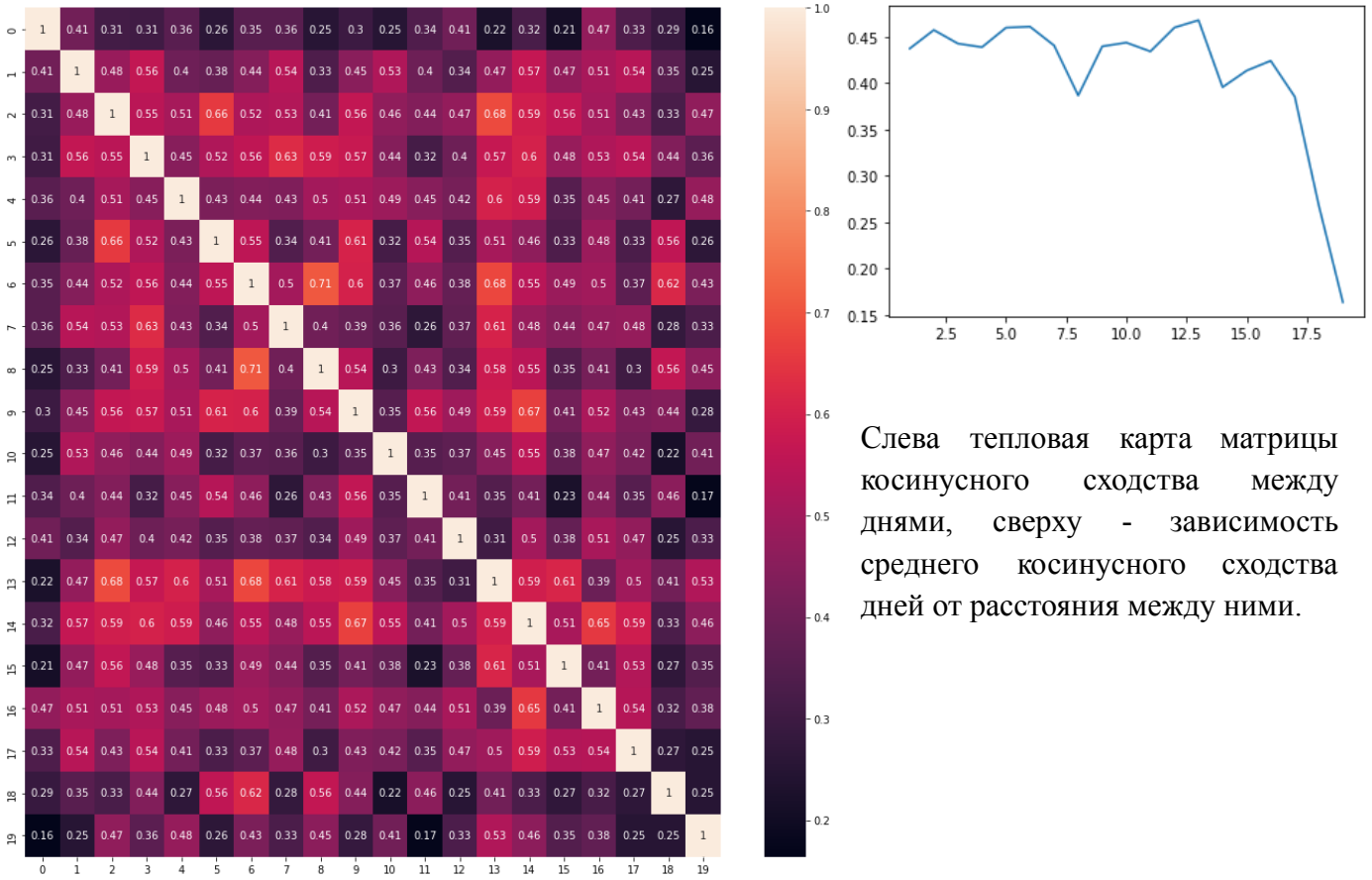
аффилиативных: больше шансов, что случайное ребро будет внутри какой-то очень тесно связанной группы.

Далее была выдвинута гипотеза о том, что схожесть аффилиативных графов за разные дни должна быть тем больше, чем дни ближе. То есть что дружественные взаимодействия обладают некоторым постоянством и завтра обезьяны будут играть друг с другом, если сегодня играли.

Для этого с помощью подхода типа bag of words (в данном случае bag of edges) для каждого наблюдаемого дня был построен вектор присутствовавших в нем аффилиативных ребер [все реализации, кроме реализации эпидемического процесса, находятся в functions.py] и затем посчитано косинусное сходство между ними. Результат отображен на следующей странице.

Визуально видно, что сходства очень много, концентрируется ли оно вокруг главной диагонали - непонятно. Чтобы это прояснить, был построен график среднего косинусного сходства между днями в зависимости от расстояния в днях между ними. Дней было всего 20. График тоже приведен ниже.

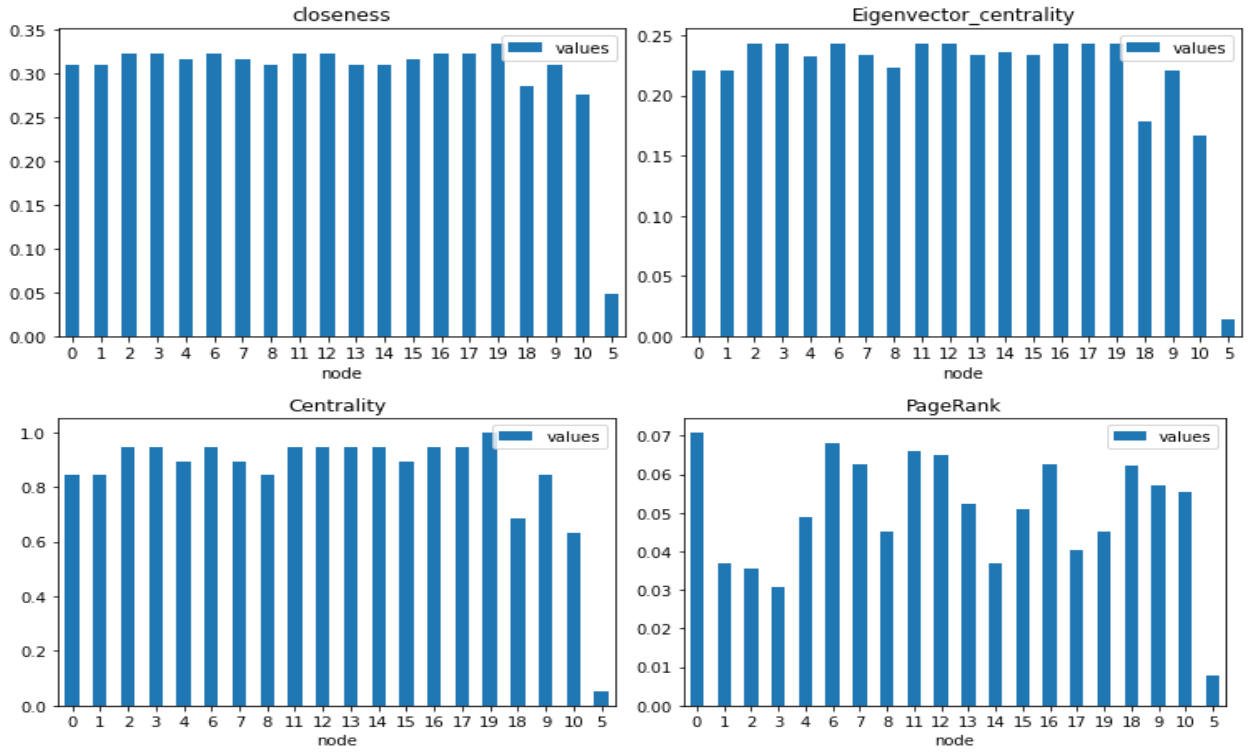
Последнюю точку не стоит принимать всерьез - есть только одна пара дней с таким расстоянием, иными словами, низкое сходство тут ничего не значит. В остальном график не выглядит убывающим, поэтому было решено дальше эту область не исследовать и связь не искать.



Слева тепловая карта матрицы косинусного сходства между днями, сверху - зависимость среднего косинусного сходства дней от расстояния между ними.

Важность вершин и community detection

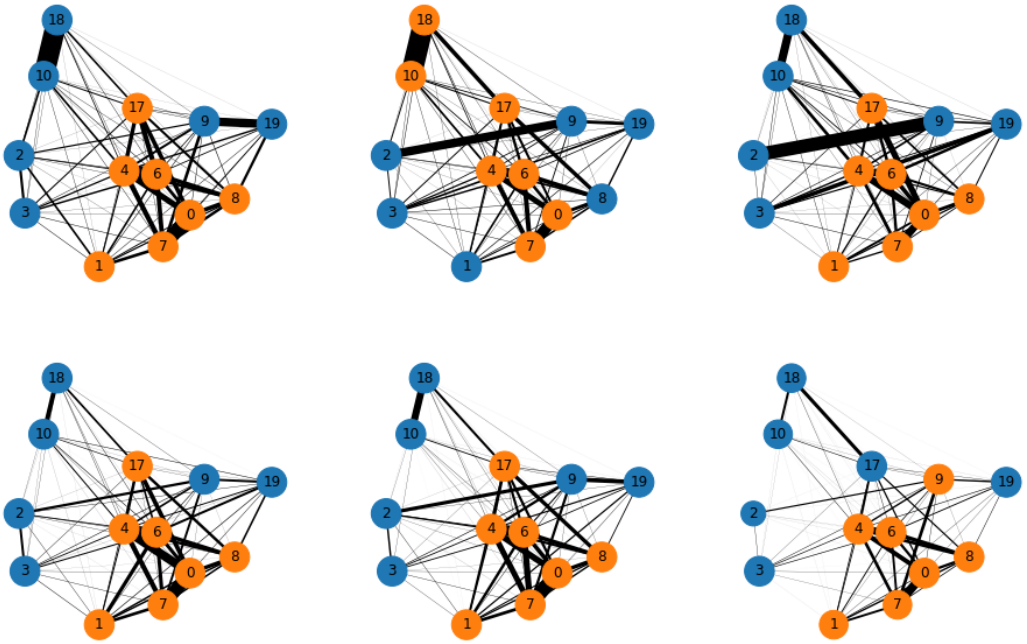
В данной части мы использовали следующие методы: degree centrality (расчет показателя центральности исходя из количества связей у узлов), eigenvector centrality (оценка центральности на основе важности узла), closeness centrality (расчет центральности на основе кратчайшего пути между нодами), pagerank (вариант eigenvector centrality). Все методы использовались с настройками по умолчанию из библиотеки Networkx. На вход почти всем алгоритмам подавался ненаправленный граф на основе всего файла OBS с признаком Duration в качестве веса, для PageRank был подготовлен направленный граф. Данные методы дали следующие оценки:



Методы degree centrality и closeness centrality ожидаемо оказались недискриптивны для данного графа в силу его малого размера и высокой связности. Результаты во многом совпадающие с результатами, полученными на предыдущем шаге, дал PageRank. Не вполне очевидным результатом оказалось то, что 18, 9, 10, 7 опередили ноды 12 и 13, а нода 0 оказалась с ними почти на одном уровне. Одно из возможных объяснений: ноды 12 и 13, по сравнению с нодами 9 и 7, чаще взаимодействуют внутри кластера из небольшой группы нод, что видно на изображении подграфа активно взаимодействующих особей. Однако это не объясняет их отличия от нод 18 и 10, взаимодействующих по-преимуществу друг с другом. В качестве одной из гипотез можно предположить, что эти ноды находятся на границах кластеров: тогда при визуализации применения алгоритмов community detection эти ноды должны чаще остальных менять принадлежность к кластерам. В отношении данных вершин это видно на ряде визуализаций (см.например ниже), однако этот вопрос требует дополнительного изучения.

Community detection в статике и динамике с помощью разных методов

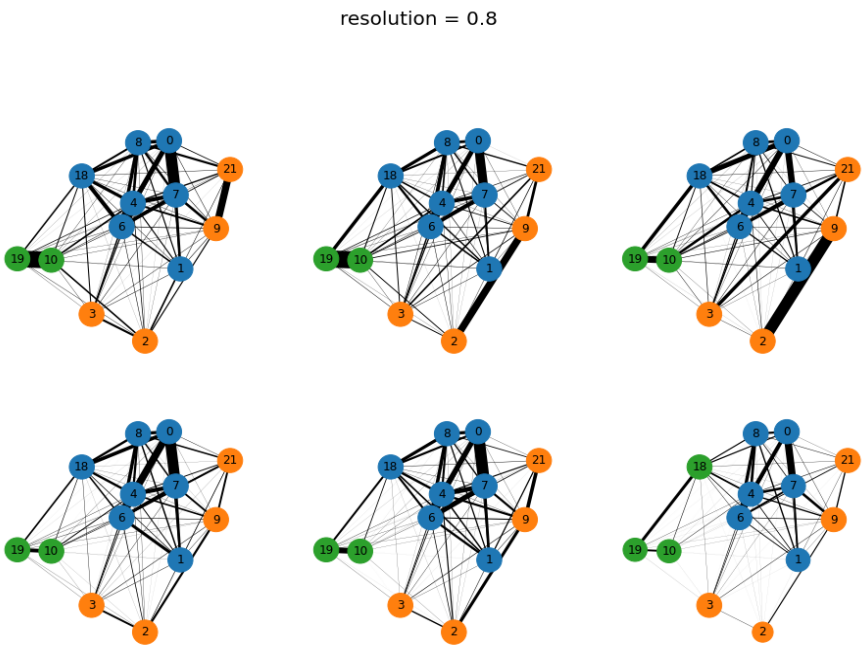
Kernighan–Lin bisection - алгоритм, разбивающий граф на две части. В разрезе 5 дней даёт достаточно стабильную структуру (датасет RFID). В разрезе 1 дня структура оказывается нестабильной.



Kernighan–Lin алгоритм в разбивке по 5 дней

У алгоритма есть существенное ограничение: он может выделить только 2 группы. На графике видно, что одну группу он выделяет по принципу сильного внутреннего взаимодействия, а другую - по остаточному принципу. Более того, для датасета OBS, где наблюдений меньше, он оказывается менее стабильным.

Greedy_modularity_communities. Этот алгоритм представляет из себя жадную оптимизацию modularity. Для данного графа это хороший выбор, т.к. количество вершин небольшое, и нет необходимости использовать более “хитрые” способы оптимизации modularity. Присутствует параметр регуляризации resolution, который штрафует за большой (>1) или маленький (<1) размер сообществ. При resolution=0.2 все схлопывается в одно сообщество. Наиболее стабильным в разрезе 5 дней граф оказывается при resolution=0.8 (RFID). При этом же параметре регуляризации алгоритм демонстрирует неплохую стабильность в разрезе 1 дня (см. ноутбук).



В тоже время для датасета OBS разбиения менее стабильны.

Эпидемии

Эпидемии моделировались на данных о сенсорном взаимодействии RFID (то есть на 13 обезьянах; их номера совпадают с номерами в анализе выше, можно сравнивать с уже имеющейся информацией). У каждого такого взаимодействия указано время в unix-формате (без длительности), время в формате datetime (просто для удобства) и пара взаимодействующих обезьян.

Параметрами эпидемии были: ее вид (SI, SIR, SIS), вероятность заражения при разовом контакте (варьировалась от 0.1 до 0.7), момент, когда заболела первая обезьяна, обезьяна, заболевшая первой и математическое ожидание времени выздоровления обезьяны (кроме эпидемий типа SI). Время выздоровления бралось из пуассоновского распределения с соответствующим мат.ожиданием. Мат. ожидание варьировалось от 1 часа до 46 часов с шагом в 5 часов.

Результатом такого перебора параметров стал датасет df_infections с 13650 строками, в каждой из которых указывались приведенные выше параметры эпидемии, а также результат: закончилась она или нет (если нет, это значит, что раньше ситуации, которая для данного вида эпидемии уже не изменится, например, нет больных для SIS и SIR или все больны/все здоровы для SI, наступил конец датасета со взаимодействиями), ее длительность, количества инфицированных / не болевших / переболевших (для SIR), а также для каждой особи ее личный статус и количество раз, которые за данную эпидемию эта особь болела и суммарное время ее болезней (которое, впрочем, должно быть примерно пропорционально количеству болезней для SIR и SIS).

Этот датасет открывает большие возможности для аналитики [и при необходимости его можно выслать, ведь считался он долго].

Очевидные выводы (много графиков в ноутбуке):

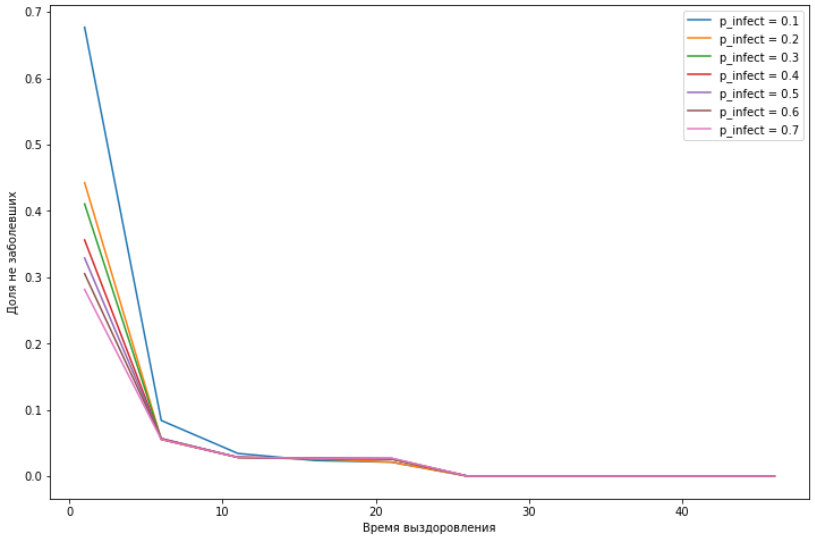
1) Стартовая вершина ни на что не влияет. Длительность эпидемии и распределение состояний особей к ее окончанию не зависят от старта. Контактов происходит очень много, поэтому начало перестает влиять.

2) С увеличением вероятности заражения при контакте падает средняя длительность эпидемии для всех трех видов эпидемий.

3) Популярные вершины заболевают первыми, непопулярные - последними.

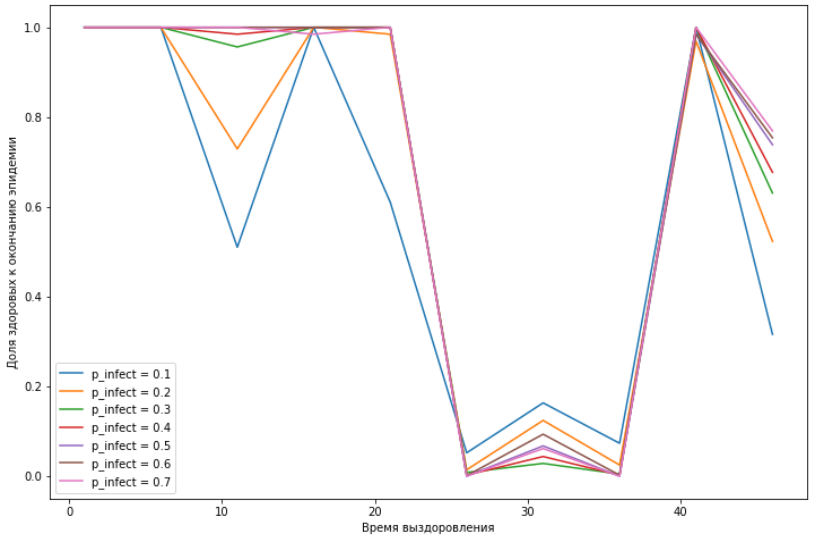
4) Эпидемии SI самые короткие (все заболели и конец), SIS - самые длинные (останавливается только если все выздоровели), SIR посередине (все либо переболеют, либо не заболеют, но переболевшие не могут заражать других - поэтому длиннее, чем SI). Здесь можно привести цифры: SI при вероятности контактной передачи болезни 0.1 в среднем длится около 9.5 часов, SIR - около 44, SIS - около 255.

Для эпидемий SIR был построен следующий график: для каждой из множества вероятностей заболеть отображалась зависимость доли так и не заболевших на момент конца эпидемии от среднего времени выздоровления инфицированного.



Он любопытен тем, что, различаясь вначале, графики очень быстро становятся очень похожи. То есть для времени выздоровления в 6 часов при имеющейся интенсивности контакта уже не очень важно, передается болезнь с вероятностью 0.1 или 0.7.

А вот с таким же графиком для SIS (там в любой момент есть только здоровые и больные, мы считаем среднюю долю здоровых на момент конца эпидемии в зависимости от времени выздоровления; концом эпидемии считается либо всеобщее здоровье, либо конец имеющихся взаимодействий) ситуация загадочна.



Средняя доля здоровых, равная единице, означает, что все эпидемии с заданным временем выздоровления (и вероятности передачи заболевания) закончились выздоровлением. Но, кажется, у нее есть объяснение! Обезьяны не взаимодействуют по ночам, последние контакты в 22 часа (и то их всего 55 на весь месяц наблюдения), а первые - около 5 утра. Таким образом, учитывая, что днем контакты очень плотные, все могут заразиться приблизительно в одно время - и, как следствие, выздороветь тоже приблизительно в одно время. И если это время придется на ночь, то новых контактов, поддерживающих болезнь, уже не будет.

К примеру, для любой вероятности контакта при времени выздоровления в 6 часов (вторая точка на графике) первая же ночь вылечит всех обезьян. Скорее всего, и дальше именно такие эффекты влияют на график. Очень любопытна точка 41: видимо, либо в первую волну заражения, либо во вторую гарантированно настанет ночь.