«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа прикладной математики и информатики
Кафедра «Интеллектуальные системы»
Бишук Антон Юрьевич

Применение активного обучения к графовым моделям на примере оценки рисков распространения эпидемии

03.03.01 — Прикладные математика и физика

Выпускная квалификационная работа бакалавра

Научный руководитель:
к.ф.-м.н.
Зухба Анастасия Викторовна

Москва
2021 г.

Содержание

Аннотация

Данная работа посвящена исследованию влияния противоэпидемиологических мер на распространение эпидемии в графовых структурах. Целью являлась разработка стратегий тестирования, вакцинации, локдауна и других возможных видов противоэпидемиологических мер. Был предложен, теоретически обоснован и использован в экспериментах метод моделирования, группирующий возможную инициализацию эпидемии по свойствам вершин. Показано, что в отличии от вакцинации и изоляции больных, локдаун может привести к увеличению математического ожидания числа заболевших. Предложен, реализован и экспериментально проверен алгоритм выбора вершин для тестирования, опирающийся на информационные критерии.

1 Введение

Эпидемия, вызванная новой коронавирусной инфекцией, привела к серьезным экономическим и психологическим последствия. Многие организации были вынуждены реорганизовать рабочие процессы таким образом, чтобы препятствовать распространению заболевания. К сожалению, многие производственные процессы невозможно осуществлять при помощи дистанционных технологий. Уменьшение количества работников, находящихся на рабочих местах, может привести к существенному снижению эффективности. С другой стороны, высокий уровень заболеваемости также может снизить эффективность производства вплоть до полной остановки технологических процессов. В связи с этим, возникла необходимость исследования различных сценариев течения эпидемии и разработки стратегий и инструментов борьбы с ней.

Методы эпидемиологии, опирающиеся только на количественные оценки числа заболевших и средние значения шансов заражения, времени выздоровления и т.д., применимы только для больших групп населения порядка города или даже страны. На маленьких группах людей поведение эпидемии может существенно отличаться от оценок «в среднем». Потому возникает необходимость анализировать индивидуальные особенности поведения людей и организации рабочих процессов. Так, например, работник столовой ежедневно общается с большим количеством людей из организации и подвергается на работе большему риску заражения, чем работник архива.

Современные технологии дают возможность отслеживать продолжительность и дистанцию общения между людьми внутри организации, что позволяет формировать динамический граф контактов.

В данной работе рассматривается распространение эпидемии на графах. Начальным этапом данного исследования была не только обработка реальных, но и генерация модельных и синтетических графов контактов, поскольку реальные данные о течении эпидемии на графе контактов, к которым получен доступ, имеют небольшой объем и не дают возможность рассмотреть различные сценарии.

Для изучения влияния противоэпидемиологических мер потребовалось моделирование различных сценариев эпидемии на графах контактов. Моделирование распространения заболевания производилось при помощи модификация модели SEIRS.

В разделе, посвященном моделированию эпидемии, были предложены методы, позволяющие учитывать время, дистанцию и другие свойства контактов при расче-

те вероятности заражения. Сформулированы и доказаны условия, дающие возможность уменьшить разброс и, таким образом, снизить число запусков моделирования эпидемии, необходимое для изучения особенностей ее поведения.

В следующих разделах рассмотрены ограничительные противоэпидемиологические меры. Введено понятие монотонности меры: ограничительная мера называется монотонной, если ее введение не увеличивает математическое ожидание количества заболевших на следующем шаге. Показано, что вакцинация и изоляция больных в отличии от локдауна, являются монотонными. Сформулирован и доказан критерий локдауна, при котором его введение может увеличить математическое ожидание количества заболевших.

Предложенные в работе стратегии противоэпидемиологических мер опираются на оценки рисков заболеть для различных вершин. Для получения информации, позволяющей рассчитывать эти риски, необходимо тестирование на наличие самого заболевания и антител. В предлагаемой модели тестирование определяет принадлежность вершины к классу. Тестирование человека – это дорогостоящее действие, что привело к постановке задачи в терминах активного обучения[1]. Активное обучение – это раздел машинного обучения, в котором алгоритм сам выбирается данные, необходимые для разметки, однако разметка каждого объекта дорогостоящая, в некотором смысле, операция – будь то время, деньги или вычислительные мощности. В связи с этим, появляется необходимость выбирать для проверки те объекты, которые наилучшим образом оптимизируют информационную метрику. В рассматриваемой задаче – объектом является человек, а его меткой – статус (болен или здоров). Получение метки означает тестирование человека на антитела, а информационной метрикой может служить, например, энтропия. Этой задаче посвящен последний раздел теоретической части данной работы.

В практической части работы приведены эксперименты, иллюстрирующие все полученные теоретические результаты, а также продемонстрировано экспериментальное сравнение алгоритмов тестирования с различными критериями.

2 Обзор литературы

Первые попытки математического моделирования эпидемиологических процессов, получившие широкую известность, появились еще в самом начале XX века[2].

Изучение эпидемий на больших популяциях позволяет моделировать этот про-

цесс «в среднем», и даже получать точные аналитические решения[3].

В зависимости от поставленной прикладной задачи возникает необходимость моделировать процесс эпидемии с разной степенью подробности. Так, например, простейшая модель SI[4] рассматривает всего два состояния: больной и здоровый. В этой модели не рассматривается формирование иммунитета: здоровый всегда может заразиться при контакте с инфекцией. Существуют модели, рассматривающие дополнительно формирование иммунитета, инкубационный период, летальные исходы и многие другие возможные состояния. Одной из таких моделей является SEIR(S)[5], модификация которой используется в данной работе.

Моделирование «в среднем» не подходит для небольших или слишком разнородных популяций. Эту проблему позволяют решить модели распространения эпидемии на графах[6], [7].

Распространение эпидемии на графе контактов можно рассматривать, например, при помощи цепи Маркова[8]. Одна из проблем анализа распространения эпидемии на графе контактов заключается в том, что на больших популяциях задача требует больших затрат по времени и необходимым объемам памяти. Частично решить эту проблему, а также сделать некоторые выводы о развитии эпидемии, позволяют приближения среднего поля[9] и теория перколяции[10]. Наиболее распространенной является задача прогнозирования течения эпидемии[11] и оценка индивидуальных рисков.

При этом приближении, подбор параметров модели существенно зависит от объема данных и решаемой прикладной задачи, что обосновывает актуальность данного исследования.

Востребованность исследования на данную тему возросла в связи с эпидемией COVID-19[12]. Однако результаты изучения распространения эпидемии на графах могут быть использованы не только для анализа заболеваний. Например, распространение слухов или автомобильного трафика можно описать схожим математическим аппаратом[13].

3 Основные понятия и подходы

Для начала определим и введем используемые основные понятия, модели и опишем их улучшения.

3.1 Общие определения

Определение 3.1. *Граф контактов – граф, у которого вершина – человек, а ребро – контакт между людьми. Ребра могут иметь веса: продолжительность контакта и т.д.*

Каждая вершина графа может находиться в конечном количестве состояний: болен, здоров, уязвим, и т.д

Определение 3.2. *Итерация (шаг моделирования) – в данной работе временной шкалой принято считать дни, а значит за один шаг (одну итерацию) принимается один день.*

Определение 3.3. *Поведение эпидемии – набор графиков, представляющих собой усредненное число вершин с определенной меткой на каждой итерации. Так, например, можно рассматривать поведение эпидемии, как среднее изменение числа заболевших с течением времени.*

Определение 3.4. *Семплирование эпидемии – многократное моделирование распространения эпидемии согласно правилам заданной математической модели.*

3.2 Математические модели распространения эпидемии

В данной работе будут использовать две математические модели распространения эпидемии: SIS будет использоваться для задачи тестирования, а более общая модель – SEIRS, для моделирования и анализа ограничительных мер.

3.2.1 SIS

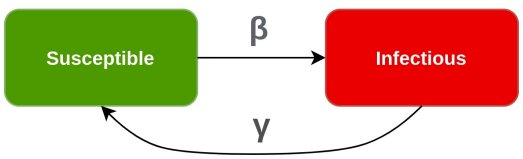

Рис. 1: Модель распространения эпидемии SIS.

$$\begin{cases} \frac{dS}{dt} = \frac{-\beta SI}{N} + \gamma I \\ \frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I \end{cases}$$

Где $S + I = N$ – вся популяция

Определение 3.5. *Параметры модели SIS:*

- β – *вероятность для восприимчивых вершин (S) заразиться;*
- γ – *скорость выздоровления, т.е. скорость перехода вершин из состояния больной (I) в состояние иммунный (R);*

3.2.2 SEIRS

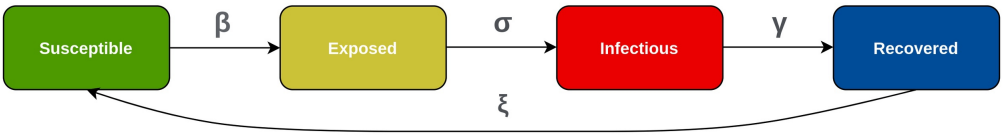

Рис. 2: Модель распространения эпидемии SEIRS.

$$\begin{cases} \frac{dS}{dt} = -\frac{\beta IS}{N} + \xi R, \\ \frac{dE}{dt} = \frac{\beta IS}{N} - \sigma E, \\ \frac{dI}{dt} = \sigma E - \gamma I, \\ \frac{dR}{dt} = \gamma I - \xi R. \end{cases}$$

Где $S + E + I + R = N$ – вся популяция

Определение 3.6. *Параметры модели SEIRS:*

- β – *вероятность для восприимчивых вершин (S) заразиться;*
- σ – *скорость перехода вершин из состояния зараженный (E) в состояние больной (I);*
- γ – *скорость выздоровления, т.е. скорость перехода вершин из состояния больной (I) в состояние иммунный (R);*
- ξ – *скорость потери иммунитета, т.е. скорость перехода вершин из состояния иммунный (R) в состояние восприимчивый (S).*

Графовые аналоги параметров SIS и SEIRS имеют тот же смысл, но не для множества вершин, а для каждой вершины в отдельности.

3.2.3 Графовый SEIRS: модернизация

Рассмотрим графовую модель SEIRS, в которой информация о структуре графа будет учитываться в переходе из S в E.

Параметры σ, γ, ξ – остаются такими же, как в классической модели, и равны константным значениям.

В уравнении модели SEIRS выделены красным те части, которые остаются неизменными по сравнению с классической моделью.

$$\begin{cases} S' = -\beta IS, \\ E' = \beta IS - \sigma E, \\ I' = \sigma E - \gamma I, \\ R' = \gamma I - \xi R. \end{cases}$$

Тогда введем следующую замену:

$$\beta IS \to \beta(\beta_{vac}, I, S, G),$$

где β_{vac} – заразность (характеристика самого заболевания), G – характеристики графа (контактные вершины, длина рёбер и т.д.).

Система рассматривается в моменты времени $T = \{t_1, t_2, \dots t_k, \dots\}$.

Рассмотрим граф $G(V, E)$ и введем следующие обозначения:

Пусть $A \in V$, тогда $S_t(A), E_t(A), I_t(A), R_t(A)$ – назовем подмножества вершин из множества вершин A, находящиеся в момент времени t в состояниях Susceptible, Exposed, Infectious, Recovered соответственно.

Определение 3.7. *Назовем $V_i = \{v_j | e_{ij} > 0\}$ – множество контактных с v_i вершин.*

Формализуем вероятности переходов из различных состояний:

$$\begin{cases} \sigma = \mathbb{P}(v_i \in I_{t+1} | v_i \in E_t), \\ \gamma = \mathbb{P}(v_i \in R_{t+1} | v_i \in I_t), \\ \sigma = \mathbb{P}(v_i \in E_{t+1} | v_i \in R_t), \\ \beta_t^i = \mathbb{P}(v_i \in E_{t+1} | v_i \in S_t). \end{cases}$$

Предложим следующий способ подсчета β:

$$\beta_t^i = |I_t(V_i)|\frac{1}{|V_i|}B_i \tag{3.1}$$

В такой постановке, структура графа учитывается явным образом через степень вершины, а B_i содержит все другие зависимости (например, время контакта, близость контакта и другие).

В частности, если рассматривать случай с двумя подструктурами контактов: дом (H) и работа (W), вероятность заболеть переписывается следующим образом:

$$\beta_t^i = \begin{cases} |I_t(V_i^H)|\frac{1}{|V_i^H|}B_i^H & \text{, на время нахождения дома} \\ |I_t(V_i^W)|\frac{1}{|V_i^W|}B_i^W & \text{, на время нахождения на работе} \end{cases} \tag{3.2}$$

Для упрощения записи введем следующее обозначение:

$$S_t := S_t(V)$$

Необходимо узнать какое число людей в среднем переходит из одного состояние в другое за один момент времени:

$$\begin{cases} \mathbb{E}(|R_{t+1}| - |R_t|) &= \mathbb{E}(\gamma|I_t| - \xi|R_t|), \\ \mathbb{E}(|I_{t+1}| - |I_t|) &= \mathbb{E}(\sigma|E_t| - \gamma|I_t|), \\ \mathbb{E}(|E_{t+1}| - |E_t|) &= \mathbb{E}(S \to E) - \mathbb{E}(\sigma|E_t|), \\ \mathbb{E}(|S_{t+1}| - |S_t|) &= \mathbb{E}(\xi|R_t|) - \mathbb{E}(S \to E). \end{cases} \tag{3.3}$$

Здесь $\mathbb{E}(S \to E)$ – математическое ожидание людей, переходящих из состояния S в состояние E в момент времени t.

Поскольку β_t^i – вероятность для i-го человека заразиться, то имеем:

$$\mathbb{E}(S \to E) = \frac{1}{2}\mathbb{E}\left(\sum_{v_i \in S_t, G^{Home}} \beta_t^i + \sum_{v_i \in S_t, G^{Work}} \beta_t^i\right)$$

Теперь если предполагать, что на k-ой итерации нам известно всё о системе, а необходимо узнать что станет с ней на $k+1$ итерации, то (3.3) переписывается в следующем виде:

$$\begin{cases} \mathbb{E}(|R_{t+1}|) - |R_t| &= \gamma|I_t| - \xi|R_t|, \\ \mathbb{E}(|I_{t+1}|) - |I_t| &= \sigma|E_t| - \gamma|I_t|, \\ \mathbb{E}(|E_{t+1}|) - |E_t| &= \sum\limits_{i:v_i \in S_t} |I_t(V_i^H)| \frac{1}{|V_i^H|} B_i^H + \sum\limits_{i:v_i \in S_t} |I_t(V_i^W)| \frac{1}{|V_i^W|} B_i^W - \sigma|E_t|, \\ \mathbb{E}(|S_{t+1}|) - |S_t| &= \xi|R_t| - \sum\limits_{i:v_i \in S_t} |I_t(V_i^H)| \frac{1}{|V_i^H|} B_i^H - \sum\limits_{i:v_i \in S_t} |I_t(V_i^W)| \frac{1}{|V_i^W|} B_i^W. \end{cases} \tag{3.4}$$

Если же мы не знаем точное состояние системы на прошлой итерации, а знаем лишь вероятности вершин принадлежать к тому или иному множеству, тогда (3.4) переписывается в следующем виде:

$$\begin{cases} \mathbb{E}(|R_{t+1}| - |R_t|) &= \mathbb{E}(\gamma|I_t| - \xi|R_t|), \\ \mathbb{E}(|I_{t+1}| - |I_t|) &= \mathbb{E}(\sigma|E_t| - \gamma|I_t|), \\ \mathbb{E}(|E_{t+1}| - |E_t|) &= -\mathbb{E}(\sigma|E_t|) + \sum\limits_{i:v_i \in V} \frac{B_i^H}{|V_i^H|} \mathbb{E}\left(|I_t(V_i^H)| \cdot \mathbb{P}(v_i \in S_t)\right) + \\ &\qquad + \sum\limits_{i:v_i \in V} \frac{B_i^W}{|V_i^W|} \mathbb{E}\left(|I_t(V_i^W)| \cdot \mathbb{P}(v_i \in S_t)\right), \\ \mathbb{E}(|S_{t+1}| - |S_t|) &= \mathbb{E}(\xi|R_t|) - \sum\limits_{i:v_i \in V} \frac{B_i^H}{|V_i^H|} \mathbb{E}\left(|I_t(V_i^H)| \cdot \mathbb{P}(v_i \in S_t)\right) - \\ &\qquad - \sum\limits_{i:v_i \in V} \frac{B_i^W}{|V_i^W|} \mathbb{E}\left(|I_t(V_i^W)| \cdot \mathbb{P}(v_i \in S_t)\right). \end{cases}$$

3.3 Разрастании эпидемии

Для того, чтобы предсказать поведение эпидемии даже в ближайшем будущем, необходимо иметь какие-либо численные оценки и критерии. Одним из таких критериев является критерий, оценивающий эпидемиологический порог по структуре динамического графа[6].

Определение 3.8. *Критический уровень заражения (эпидемиологический порог)* λ_c *– значение уровня заражения, такое, что при значении* λ *меньше его, эпидемия затухает, а при значении* λ *больше* λ_c*, эпидемия разрастается.*

Определение 3.9. *Средняя связность* $\langle k \rangle$ *– величина, равная* $\sum\limits_k kP(k)$*, где* $P(k)$ *– распределение связности в графе.*

Теорема 3.1. *В модели SIR критическое значение уровня заражения выражается формулой:*

$$\lambda_c = \frac{\langle k \rangle}{\langle k^2 \rangle}$$

Данная теорема показывает что возможно построить критерий роста больных при эпидемии, имея только информацию о структуре графа, что натолкнуло нас на формулировку критерия «плохого» локдауна.

3.4 Активное обучение и теория информации

Активное обучение - это подраздел машинного обучения. Ключевая гипотеза состоит в том, что если алгоритму для обучения разрешено выбирать данные, на основе которых он будет обучаться, то он будет работать с таким же качеством, но при меньшей обучающий выборке. Для того, чтобы любой контролируемый обучающийся алгоритм работал хорошо, его, часто, нужно обучать на сотнях или даже тысячах помеченных объектах. Иногда метки объектов предоставляются за небольшую плату или бесплатно. Например, отметка «спам», которую пользователь отмечает на нежелательных сообщениях на электронной почте, или пятизвездочная оценка, которую он может дать фильмам в онлайн кинотеатрах.

Алгоритмы машинного обучения используют эти флаги и рейтинги, чтобы лучше фильтровать нежелательную почту и предлагать фильмы, которые могут понравиться конкретному человеку. Эти метки достаются без дополнительных затрат, но это нетипичная ситуация. Во многих других задачах обучения с учителем, получение размеченных объектов очень сложно – занимает много времени или дорого. Например, разметка данных в задаче распознания речи, классификации изображений, детекции объектов и многих других, требует определенной квалификации для того, чтобы полученный набор обучающих данных оказался релевантным.

Все сценарии активного обучения включают оценку информативности неразмеченных объектов, которые могут либо каждый раз генерироваться случайным образом, либо браться из заданного распределения. В литературе было предложено множество способов формулировать такие стратегии запросов, например, least confident[14], [15] , margin sampling[16] и стратегии основанные на энтропии[17]. Последний подход формально описывающийся как:

$$x_H^* = \arg\max_x \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x)$$

и будет использоваться в данной работе.

4 Моделирование и анализ эпидемии

Реальные данные о протекании эпидемии в небольших системах на данный момент либо немногочисленны, либо закрыты. Кроме того, одна из реализаций эпидемии не даёт комплексного представления о распространении эпидемии в среднем. Из-за этого возникла задача моделирование эпидемии согласно заданным математическим моделям. Данный подход не только позволяет получать множество реализаций эпидемии на заданных графах, но и моделировать ограничительные меры, а после сравнивать течения эпидемии.

4.1 Моделирование эпидемии на графах

Моделирование эпидемии на графе происходит путём разыгрывания случайной величины, соответствующей переходу из текущего состояния. Так, если вершина была восприимчивой (S), то вероятность стать заразившейся (E) получается агрегацией рисков от каждой смежной с ней вершины. Если же вершина заразившаяся (E), то она с вероятностью σ перейдет в состояние больной (I), а с вероятностью $1 - \sigma$ останется в состоянии (E), аналогично для перехода из состояния I в состояние R и вероятностью γ и переходом из состояния R в состояние S и вероятностью ξ.

Если рассматриваемая система не является закрытой, то вводится параметр заражения от внешней среды P и тогда каждая вершина дополнительно с вероятностью P может изменить своё состояние с S на E.

В качестве основы алгоритма моделирования использовалась библиотека SEIRS+[18], в которой были реализованы все основные свойства модели SEIRS, описанные выше. В процессе работы был расширен функционал модели до возможности тонкой настройки протекания эпидемии, возможности использовать произвольную функцию пересчета рисков (в том числе на основе времени контактов), возможности моделировать ограничительные меры.

4.2 Теорема об эквивалентности

С теоретической точки зрения, состояние системы в конкретный момент времени удобно рассматривать как состояние марковской цепи.

Определение 4.1. *Назовем вектор x – вектором рисков, представляющий из себя вектор длины 2^n, где n – число вершин в графе, строящийся по следующему правилу: на первом месте стоит вероятность того, что никто не болеет, далее вероятность, что болеет только первая вершина, болеют первые две вершины, . . ., болеют все вершины.*

Определение 4.2. *Назовем матрицу A – матрицей распространения эпидемии, представляющую из себя матрицу размера $2^n \times 2^n$, строящуюся по определенным правилам: в строках находятся состояния системы на текущей итерации, в столбцах состояния, в которые возможно перейти, в ячейках матрицы – вероятность такого перехода.*

Замечание 4.1. A^k - представление эпидемии на k-ом шаге. Если взять за x - вектор риска в начальный момент времени, то $x^T A^k$ - вектор рисков на k-ой итерации.

Что формально показывает, что течение эпидемии это переход из одного состояния в другое в марковской цепи.

Определение 4.3. *Назовем агрегацией матрицы A операцию $A^k_{2^n \times 2^n} \to \overline{A}^k_{2^n \times (n+1)}$, осуществляемую умножением матрицы A на матрицу F такого вида, что в получившейся матрице $\overline{A}^k$ столбцы формируются путем сложения столбцов матрицы A соответствующих одному и тому же числу больных.*

Определение 4.4. *Если в матрице A^k из одной строки можно получить другую путем перенумерации вершин графа, то такие состояния будем называть эквивалентными в глобальном смысле.*

Определение 4.5. *С точки зрения числа больных удобно рассматривать матрицу $\overline{A}^k$. Тогда строки, которые совпадают в агрегированной матрице будем называть эквивалентными в локальном смысле.*

Определение 4.6. *Будем называть размером окрестности степени k вершины v – число вершин, до которых можно добраться следуя по ребрам исходного графа не более чем за k переходов.*

Замечание 4.2. Если первые окрестности вершин v_1 и v_2 совпадут, то строки матрицы $\overline{A}$, соответствующие случаям, когда болеет только v_1 и только v_2 так же совпадут, т.е. будут эквивалентны в локальном смысле.

Теорема 4.1 (Бишук, 2021). *Пусть дан граф $G(V, E)$, в нем n вершин заражено и $n - V$ здоровых, причем окрестности больных не пересекаются. Тогда число заболевших на следующей итерации монотонно от величины окрестности для всех больных вершин.*

Доказательство.

Рассмотрим зараженную вершину (v_j) со степенью k. Рассмотрим подграф исходного графа, являющийся окрестностью данной вершины.

Поскольку нам известна структура графа, то мы можем составить матрицу перехода системы A для подграфа.

В строке, соответствующей нашей ситуации (v_j – больная, а все остальные здоровые), число ненулевых элементов равно $C_{k+1}^1 + C_{k+1}^2 + ... + C_{k+1}^{k+1}$.

Агрегируем матрицу A и получим $\overline{A}$.

Тогда в матрице $\overline{A}$ в строке j, в первом столбце будет стоять одно число, означающее вероятность того, что никто не заболел, а текущий человек выздоровел, во втором столбце будет сумма из C_{k+1}^1 чисел, которая означает вероятность того, что заболел один человек или v_j – не выздоровел, т.д. и в $(k+1)$-ом столбце будет стоять вероятность того, что заболело k человек и v_j не выздоровел.

Поскольку получено распределение числа больных на следующий шаг, то можем посчитать математическое ожидание числа заболевших на следующей итерации. Для этого умножим строку матрицы $\overline{A}$ на вектор $w = [0, 1, 2, ..., |V|]$ (будем считать $w_i = i - 1$), представляющий из себя вектор из числа заболевших.

Обозначим число заболевших на 1-ой итерации за I_1. Тогда:

$$\mathbb{E}(I_1) = \sum_{i=1}^{k+1}(i-1) \cdot \sum_{l=1}^{C_{k+1}^i} p_{i,l} \tag{4.1}$$

Здесь $p_{i,l}$ – вероятности перехода в матрице A в группе i (группа формируется по количеству заболевших) по номеру ненулевого элемента l.

Как видно из выражения – математическое ожидание числа заболевших монотонно по величине окрестности.

Величина 4.1 рассчитывается для каждой вершины отдельно, а затем складывается. Сумма монотонна по размеру окрестности каждой из заболевших вершин.

Таким образом утверждение теоремы доказано.

■

Замечание 4.3. Случай пересекающихся окрестностей является более громоздким, т.к. требует аккуратности в определении монотонности и рассмотрении случаев, но доказывается аналогично.

4.3 Ограничительные меры и их моделирование

Для борьбы с распространением эпидемии используется ряд противоэпидемиологических мер:

- Вакцинация,
- Изоляция больных,
- Введение lockdown.

Далее будет показано, что первые два способа ограничения эпидемии в худшем случае могут не помочь в сдерживании эпидемии, но локдаун – ограничение, которое может отрицательно повлиять на поведение эпидемии.

Моделирование вакцинации происходит путём замены статуса человека с S на R. Поскольку такая модель отличается от реальной жизни, где иммунитет появляется только через несколько недель, а до того момента, человек становится более подвержен заражению, будем считать, что данная операция проводится с многими вершинами, но успешно иммунитет приобретают только выбранные. Это увеличивает цену данной операции, что в достаточной степени приближает реальную ситуацию.

Изоляции больных моделируется путём удаления вершины из графа и всех инцидентных с ней рёбер, либо заменой метки вершины с I на S, подразумевая, что у компании есть резервные сотрудники, которые могут заменить больного.

Замечание 4.4. Изоляция вершин происходит после тестировании, т.е. подразумевает ранжирование людей по вероятности того, какой человек вероятнее всего болен.

Моделирование lockdown происходит путём замены имеющегося графа контактов «работы» на граф контактов «дом», состоящий из клик небольшого размера.

Лемма 4.1. *Тестирование и изоляция больных не увеличивает число заболевших.*

Доказательство.

- Пусть была протестирована вершина и она оказалась здоровой. Это это значит, что вершину изолировать не нужно, поскольку она не сможет никого заразить, но может заразиться, что увеличит число заболевших, однако это бы произошло с той же вероятность, даже если бы она не была протестирована.
- Пусть мы проверили вершину (пусть v_i) и она оказалась больной. Если эта вершина не будет изолирована, то верхней оценкой прироста числа заболевших от этой вершины будет $\sum\limits_{v_j \in V_i} \beta$. А если эта вершина будет изолирована, то верхней оценкой будет 0.

■

Лемма 4.2. *Вакцинация не увеличивает число заболевших.*

Доказательство.

- Пусть мы выбрали вершину и провакцинировали её. Если она в будущем будет изолированной, то вакцинация никак не повлияет на прирост числа зараженных.
- Пусть вершина не изолированная и имеет больных соседей, тогда вакцинация гарантированно не увеличит число заболевших на следующих итерациях поскольку имунная вершина не может ни заразить, ни заразиться.

■

4.4 Позитивное и негативное влияние локдауна

Замечание 4.5. Пусть β в (3.2) линейно зависит от времени проведенного в графе дома и работы соответственно через B_i^H и B_i^W:

$$\beta_t^i = |I_t(V_i)| \frac{1}{|V_i|} \tilde{B}_i \cdot \tau$$

Лемма 4.3 (Бишук, 2021). *В случае линейной зависимости заражения от времени, шанс заразиться для вершин v_i растет при локдауне тогда и только тогда, когда*

$$\frac{I_k(V_i^H)\tilde{B}_i^H}{|V_i^H|} \geqslant \frac{I_k(V_i^W)\tilde{B}_i^W}{|V_i^W|}$$

Доказательство.

Пусть τ – время, в течение которого рассматривается распространения эпидемии в сутках. Например если считается, что человек может заразиться в любой момент времени, то τ можно считать равных 24 часа.

Вероятность заразиться, когда локдаун не введен:

$$\beta_k^i = |I_k(V_i^H)|\frac{\tau^H\tilde{B}_i^H}{|V_i^H|} + |I_k(V_i^W)|\frac{\tau^W\tilde{B}_i^W}{|V_i^W|}, \quad \tau = \tau^H + \tau^W$$

Вероятность заразиться после введения локдауна:

$$\tilde{\beta}_k^i = |I_k(V_i^H)|\frac{\tau\tilde{B}_i^H}{|V_i^H|}$$

Раз шанс заразиться должен увеличиться, то это значит:

$$\tilde{\beta}_k^i \geqslant \beta_k^i$$

Или

$$|I_k(V_i^H)|\frac{\tau\tilde{B}_i^H}{|V_i^H|} \geqslant |I_k(V_i^H)|\frac{\tau^H\tilde{B}_i^H}{|V_i^H|} + |I_k(V_i^W)|\frac{(\tau-\tau^H)\tilde{B}_i^W}{|V_i^W|}$$

$$|I_k(V_i^W)|\frac{(\tau-\tau^H)\tilde{B}_i^W}{|V_i^W|} \leqslant |I_k(V_i^H)|\frac{(\tau-\tau^H)\tilde{B}_i^H}{|V_i^H|}$$

что завершает доказательство. ∎

Следствие 4.1.1. *Если есть информация только о том, какой процент населения болен (обозначим это значение как q), то можно выделить следующий вид условия леммы:*

$$\frac{\tilde{B}_i^H}{\tilde{B}_i^W} \geqslant \frac{|I_k(V_i^W)|}{|V_i^W|}\cdot\frac{|V_i^H|}{|I_k(V_i^H)|} =$$

$$= \frac{\mathbb{P}(\text{У вершины } v_i \text{ есть больные вершины в соседях графа } G^H)}{\mathbb{P}(\text{У вершины } v_i \text{ есть больные вершины в соседях графа } G^W)} \approx \frac{1-(1-q)^{|V_i^H|}}{1-(1-q)^{|V_i^W|}}$$

Теорема 4.2 (Бишук, 2021). *Локдаун уменьшает математическое ожидание числа заболевших тогда и только тогда, когда:*

$$\mathbb{E}(\tilde{I}_{k+1} - I_{k+1}) = \sum_{i:v_i \in I_k} (\tilde{\beta}_i^k - \beta_i^k) = \sum_{i=1}^{|V|} [v_i \in S_k] \frac{B_i^H |V_i^W| |I_k(V_i^H)| - B_i^W |V_i^H| |I_k(V_i^W)|}{|V_i^H| \cdot |V_i^W|} > 0$$

4.5 Энтропия и тестирование людей

Тестирование людей не влияет на поведение эпидемии, однако оно позволяет уточнить информацию о системе и на основе актуализированной информации сделать вывод о состоянии конкретных людей. В данном разделе будет рассматриваться подход модели SIS.

4.5.1 Энтропия графа

Рассмотрим граф как случайную систему, в вершинах, которой записаны вероятности того, что она имеет метку 1 (вершина больна). Иными словами чем меньше вероятность находится в вершине, тем выше вероятность того, что в вершина имеет метку 0.

Тогда под энтропией будем понимать следующую величину:

$$\mathbb{H}(\{x_i\}_{i=1}^n) = -\left(\sum_{i=1}^n p_i \log(p_i) + \sum_{i=1}^n (1-p_i)\log(1-p_i)\right) \tag{4.2}$$

Максимум достигается в точке $x_{opt} = \frac{1}{2}$, это значит, что проверка человека, вероятность которого ближе всего к x_{opt} максимально уменьшит энтропию системы, получаемую от одного слагаемого.

4.5.2 Случай знания прошлого

Известна вся информация о системе на прошлой итерации.

1. Если i-ая вершина была здорова и не имела контактов с больными, то $p_i^{t-1} = p_i^t = 0$.

2. Если i-ая вершина была больна, то $p_i^t = 1 - \gamma$.

3. Если i-ая вершина была здорова, но контактировала с $\{x_{i_k}\}_{i_k}^K$ больными вершинами, то $p_i = f(\beta, K)$ (где K– число больных соседей вершины с номером i) (или в общем случае $f(\{\beta_l\}, \{\tau_l\})$).

   Где $f(\beta, K)$ можно выбрать следующими:

1) β;

2) $\beta + (1-\beta)\sum\limits_{l=1}^{K-1}(\frac{1}{2})^l$;

3) $\beta + (1-\beta)e^{-c/(K-1)}$;

4) В общем случае: $\beta + (1-\beta)\sum\limits_{l=1}^{K-1} g(l)$,

$$g(l): \begin{cases} \sum\limits_{l=1}^{K-1} g(l) \to 1 & ,K \to \infty \\ g(m) > g(m+1) & m \in \mathbb{N} \end{cases}$$

4.5.3 Случай распределения вероятностей

Пусть неизвестны метки вершин на прошлой итерации, но известно распределение вероятностей $\{p_i^{t-1}\}_{i=1}^n$. Есть два подхода к определению β:

1. β – это вероятность заразиться от больного, то есть β – верхняя оценка вероятности заразиться;

2. β – вероятность заразиться, если в окружении есть только один заболевший.

Тогда определим пересчет вероятности быть больным:

1. Если $\exists i'$: $p_{i'}^{t-1} = 0$, $\forall j : \exists (i'j) : p_j^{t-1} = 0 \hookrightarrow p_{i'}^t = 0$

2. Если $\exists i'$: $p_{i'}^{t-1} = 1$, $\hookrightarrow p_{i'}^t = 1 - \gamma$

3. Иначе вероятность пересчитывается как

$$p_i^t = p_i^{t-1}(1-\gamma) + (1-p_i^{t-1}) f(\beta, \{p_{i_k}\}), \tag{4.3}$$

где $f(\beta, \{p_{i_k}\}_{k=1}^K)$ можно выбрать следующими:

1) Наивный вариант:

$$\frac{\beta}{K}\sum_{k=1}^{K} p_{i_k}; \tag{4.4}$$

2) Определение через вернхюю границу:

$$f(\beta, \{p_{i_k}\}_{k=1}^K) = \beta e^{-c/\sum\limits_k p_{i_k}} \tag{4.5}$$

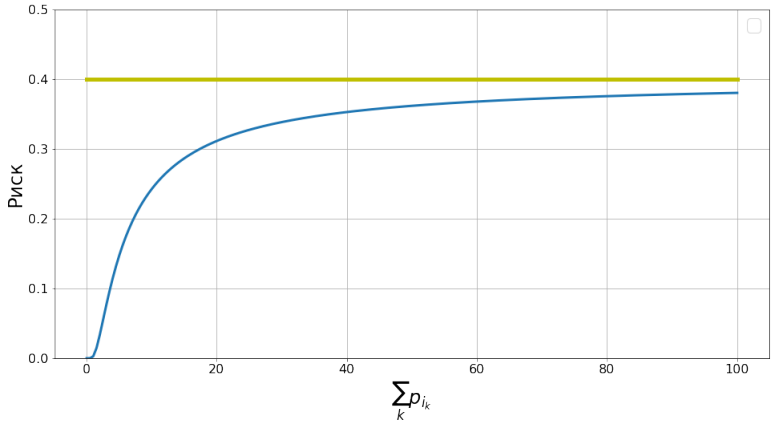

Рис. 3: Функция пересчета рисков (4.5), $\beta = 0.4$, $c = 5$. Ограничение риска через β является недостаточно гибким для настройки пересчета заразности.

3) Определение через верхнюю границу в общем случае:

$$f(\beta, \{p_{i_k}\}_{k=1}^{K}) : \begin{cases} f(\beta, \{p_{i_k}\}_{k=1}^{K-1}) < f(\beta, \{p_{i_k}\}_{k=1}^{K}) \\ f(\beta, \{p_{i_k}\}_{k=1}^{K}) \to \beta \\ \forall n, \forall p_{i_k} \hookrightarrow f(\beta, \{p_{i_k}\}_{k=1}^{K}) < \beta \end{cases}$$

4) Определение через одного больного №1:

$$f(\beta, \{p_{i_k}\}_{k=1}^{K}) = e^{\frac{\ln(\beta)}{\sum_k p_{i_k}}} \tag{4.6}$$

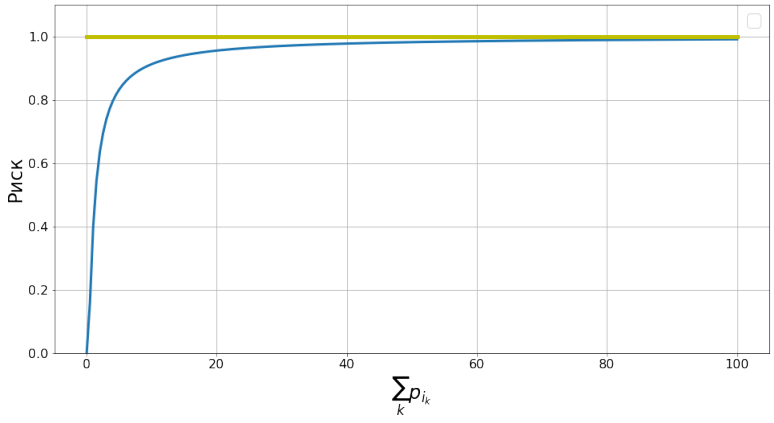

Рис. 4: Функция пересчета рисков (4.6), $\beta = 0.4$. Проблема в потери степени свободы, способной задавать поведение выпуклости.

5) Определение через одного больного №2:

$$f(\beta, \{p_{i_k}\}_{k=1}^K) : \begin{cases} \beta e^{c\left(1-\frac{1}{\sum\limits_k p_{i_k}}\right)} & \sum\limits_k p_{i_k} < 1 \\ \beta + (1-\beta)e^{c\left(\frac{-1}{\sum\limits_k p_{i_k}}\right)} & \sum\limits_k p_{i_k} \geqslant 1 \end{cases} \tag{4.7}$$

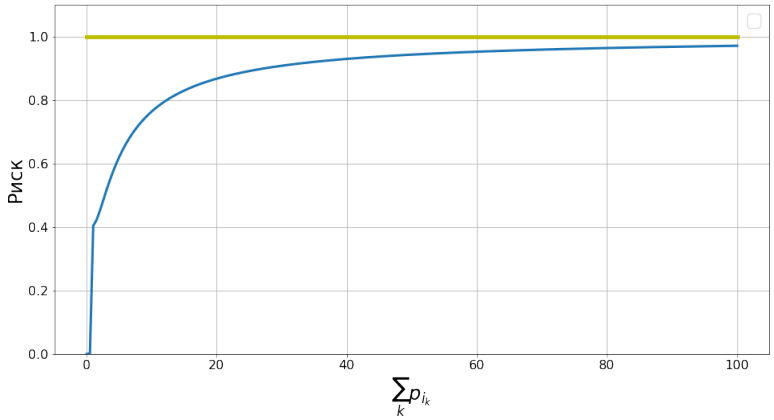

Рис. 5: Функция пересчета рисков (4.7), $\beta = 0.4$, $c = 5$. Несмотря на то, что функция непрерывная, она не является гладкой.

6) Определение через одного больного №3:

$$f(\beta, \{p_{i_k}\}_{k=1}^K) = \left(\frac{2 \cdot \operatorname{arctg}\left(\left(\sum\limits_k p_{i_k}\right)^c\right)}{\pi} \right)^{-\log_2 \beta} \tag{4.8}$$

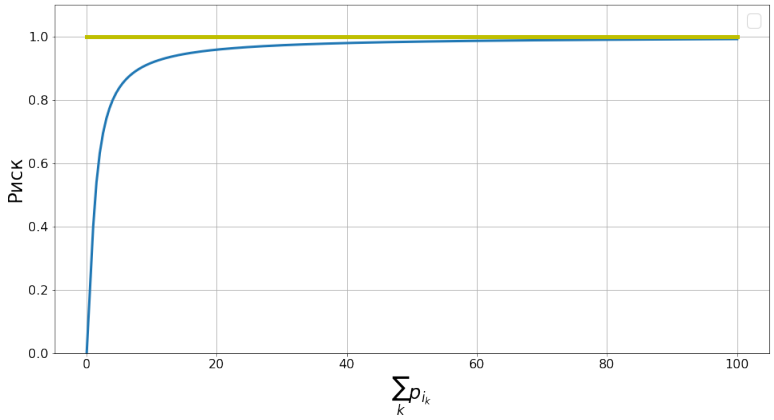

Рис. 6: Функция пересчета рисков (4.8), $\beta = 0.4$, $c = 1$.

7) Определение через одного больного в общем случае:

$$f(\beta, \{p_{i_k}\}_{k=1}^{K}) : \begin{cases} f(\beta, \{p_{i_k}\}_{k=1}^{K-1}) \leqslant f(\beta, \{p_{i_k}\}_{k=1}^{K}) & ,\forall K, \beta, p_{i_k} \\ f(\beta, \{p_{i_k}\}_{k=1}^{K}) \to 1 & ,\sum\limits_{k} p_{i_k} \to \infty \\ f(\beta, \{p_{i_k}\}_{k=1}^{K}) < 1 & ,\forall K, \beta, p_{i_k} \\ f(\beta, \{p_{i_k}\}_{k=1}^{K} = 1) = \beta & ,\forall \beta \\ f(\beta, \{p_{i_k}\}_{k=1}^{K} = 0) = 0 & ,\forall \beta \\ f(\beta, \{p_{i_k}\}_{k=1}^{K}) - \text{непрерывная} & ,\forall K, \beta, p_{i_k} \end{cases}$$

Замечание 4.6. Если известно распределение вероятностей на прошлой итерации $\{p_i^{t-1}\}_{i=1}^{n}$ и у части вершин известна их точная метка, то по построению модели, метки можно воспринимать как вероятности, поэтому для этого случая верны все формулы из описываемого пункта.

4.5.4 Алгоритм выбора

Теперь будем рассматривать подход к уменьшению энтропии и, в первую очередь, обозначим использование (4.6), (4.7) или (4.8) как функции пересчета рисков.

Определение 4.7. *Локальной энтропией вершины v_i на шаге t будем называть величину:*

$$h_i^t = H_i^t + \sum_k H_{i_k}^t$$

Будет производиться тестирование вершины и анализироваться изменение энтропии системы на следующем шаге, а также анализироваться качество приближения вероятностями меток.

Сам алгоритм представляет из себя систему пересчета рисков на основе рисков соседей с возможностью внесения информации посредством тестирования и выдачи ранжирования вершин по заданному критерию.

Алгоритм 4.1. Алгоритм выбора вершин на основе энтропийного критерия

Вход: Массив графов [G_1(V(p),E),G_2(V,E), ..., G_τ(V(p),E)], одна из функций пересчета (4.6), (4.7), (4.8), β, c, γ.

Выход: rank – массив отранжированных вершин для каждого шага

для $i = 1, \ldots, |V|$
  для $j = 1, \ldots, |V_i|$
    G_1[i][p_{neigh}][j] $\leftarrow$ G_1[j][p]
для $t = 2, \ldots, |\tau|$
  для $i = 1, \ldots, |V|$
    G_t[i][H_i] $\leftarrow$ $H_i^t(G_t)$
  для $i = 1, \ldots, |V|$
    G_t[i][h_i] $\leftarrow$ $h_i^t(G_t)$; G_t[i][h_i] $\leftarrow$ $h_i^t(\tilde{G}_t^i)$; G_t[i][h_i] $\leftarrow$ $h_i^t(\tilde{\tilde{G}}_t^i)$
    $\tilde{G}_t^i$ ($\tilde{\tilde{G}}_t^i$) – граф полученный после замены вероятности i-ой вершины графа G_{t-1} на 1 (0) и пересчета вероятностей на следующий шаг.
  если Выдать ранжирование **то**
    Вершины сортируются по убыванию $E\Delta h_i^t(G_t, i)$ и выдаются.
  если Протестировать k-ую вершину **то**
    если k-ая – вершина больна **то**
      G_t[i][h_i] $\leftarrow$ G_t[i][$\tilde{h}_i$]; G_t[i][p] $\leftarrow$ 1; G_t[i][H_i] $\leftarrow$ 0
    иначе
      G_t[i][h_i] $\leftarrow$ G_t[i][$\tilde{\tilde{h}}_i$]; G_t[i][p] $\leftarrow$ 0; G_t[i][H_i] $\leftarrow$ 1
  для $i = 1, \ldots, |V|$
    для $j = 1, \ldots, |V_i|$
      G_{t+1}[i][p_{neigh}][j] $\leftarrow$ G_t[j][p]
    G_{t+1}[i][p] = f_{one}(G, i, β, γ, c)

Алгоритм 4.2. h_i^t

Вход: G, k

h $\leftarrow$ G[k][H_i]
для $j = 1, \ldots, |V_k|$
  h += G[j][H_i]

Выход: h

Алгоритм 4.3. $E\Delta h_i^t$

Вход: G, k

Выход: G[k][h_i] - (G[k][p]·G[k][$\tilde{h}_i$] + (1 - G[k][p])·G[k][$\tilde{h}_i$])

Алгоритм 4.4. f_{one}

Вход: G, i, β, γ, c

Выход: $G[\mathrm{i}][p](1-\gamma) + (1 - G[\mathrm{i}][p]) \left(\frac{2 \cdot \mathrm{arctg}\left(\left(\sum\limits_k G[i_k][p] \right)^c \right)}{\pi} \right)^{-\log_2 \beta}$

4.5.5 Алгоритм аккумуляции энтропии

Задача аккумуляции энтропии формулируется следующим образом: необходимо найти такую вершину x_i, изменение вероятности p_i^{t-1} которой (путем тестирования) максимально приблизит p_j^t какой-то вершины j к значению $\frac{1}{2}$. Более формально:

Пусть p_i^t – вероятность того, что в i-ой вершине находится метка 1 до того, как кто-то был протестирован, а $\tilde{p}_i^t$ – вероятность того, что в i-ой вершине находится метка 1 после того, как кто-то из его соседей был протестирован.
Тогда необходимо найти такую вершину $x_{i'}$: $\forall i \hookrightarrow |p_i^t - \frac{1}{2}| > |\tilde{p}_{i'}^t - \frac{1}{2}|$ и для $\forall i \neq i' \hookrightarrow |\tilde{p}_i^t - \frac{1}{2}| > |\tilde{p}_{i'}^t - \frac{1}{2}|$

Замечание 4.7. Теперь легко заметить, что для наивного пересчета вероятностей (4.4) выполнено $\mathbb{E}_j \tilde{p}_i^t = p_i^t, \ \forall j$, что значит в такой формулировке нет смысла проверять кого-то с прошлой итерации.

Далее будем использовать функцию пересчета (4.5).

Тогда:

$$\Delta p_i^t = \begin{cases} (1-p_i^{t-1})\beta \left(-e^{-c/\sum\limits_k p_{i_k}^{t-1}} + e^{-c/(\sum\limits_k p_{i_k}^{t-1} + 1 - p_j^{t-1})} \right) & \text{, с вероятностью } p_j^{t-1} \\ (1-p_i^{t-1})\beta \left(-e^{-c/\sum\limits_k p_{i_k}^{t-1}} + e^{-c/(\sum\limits_k p_{i_k}^{t-1} - p_j^{t-1})} \right) & \text{, с вероятностью } 1 - p_j^{t-1} \end{cases}$$

А значит:

$$\mathbb{E}_j \Delta p_i^t = (1-p_i^{t-1})\beta \left(-e^{-c/\sum\limits_k p_{i_k}^{t-1}} + (1-p_j^{t-1}) e^{-c/(\sum\limits_k p_{i_k}^{t-1} - p_j^{t-1})} + p_j^{t-1} e^{-c/(\sum\limits_k p_{i_k}^{t-1} + 1 - p_j^{t-1})} \right)$$

И тогда: $\tilde{p}_i^t = p_i^t + \mathbb{E}_j \Delta p_i^t$

Тогда необходимо решить следующую оптимизационную задачу:

$$\left| \tilde{p}_i^t - \frac{1}{2} \right| \to \min_{i,j}$$

Или если перейти к гладкой версии:

$$\left(\tilde{p}_i^t - \frac{1}{2} \right)^2 \to \min_{i,j}$$

Для простоты записи будем считать, что $c = 1$, после чего функция для оптимизации приобретает следующий вид:

$$F = \left(\frac{1}{2} - (1-\gamma) p_i^{t-1} - \beta (1 - p_i^{t-1}) \left(q_j^{t-1} e^{\frac{-1}{\sum_k p_{i_k}^{t-1} + 1 - q_j^{t-1}}} + (1 - q_j^{t-1}) e^{\frac{-1}{\sum_k p_{i_k}^{t-1} - q_j^{t-1}}} \right) \right)^2$$

Оптимум функции F необходимо находить любым методом численной оптимизации[19]. Тогда получим решение ${q_{opt}}_j^{t-1} = Q(\sum_{i_k} p_{i_k}^{t-1}, p_i^{t-1}, \gamma, \beta)$.

Поскольку данная функция меняет знак второй производной, то минимум достигается либо в оптимуме, либо на границе, т.е. необходимо проверить три точки.

Иными словами, оптимум либо в $q_j^{t-1} = 0$, либо $q_j^{t-1} = {q_{opt}}_j^{t-1}$, либо $q_j^{t-1} = \min\left(1, \sum_k p_{i_k}^{t-1}\right)$

Тогда сам алгоритм определения лучшего кандидата для соседа (Алгоритм 4.7)

Алгоритм 4.5. f_{up}

Вход: G, i, β, γ, c

Выход: G[i][p]$(1-\gamma)$ + (1 - G[i][p])$\beta e^{-c/\sum_k G[i_k][p]}$

Алгоритм 4.6. F

Вход: G, i, β, γ, c

Выход: $\left(\frac{1}{2} - (1-\gamma) p_i^{t-1} - \beta (1 - p_i^{t-1}) (x \exp\left(\frac{-1}{\sum_k G[i_k][p] + 1 - x}\right) + (1-x) \exp\left(\frac{-1}{\sum_k G[i_k][p] - x}\right)) \right)^2$

Алгоритм 4.7. Алгоритм аккумуляции энтропии (сложность $O(V+E)$)

Вход: Массив графов [G_1(V(p),E),G_2(V,E), ..., G_τ(V(p),E)], функция $f(\beta, \{p_{i_k}\})$ пересчета рисков 4.5.

Выход: rank – массив пар (целевая вершина, лучший сосед), отранжированных по величине изменения сконцентрированной в целевой вершине энтропии.

1: **для** $t = 2, \ldots, |\tau|$
2: **для** $i = 1, \ldots, |V|$
3: G_t[i][p] $\leftarrow f_{up}(G_{t-1}$, i, β, γ, $c)$
4: G_t[i][p_{best}] $\leftarrow$ SOLUTION($F'_x = 0$) OR 0 OR MIN(1, $\sum\limits_k G[i_k][p]$)
5: M = $|G_t$[i][p_{best}] - $G_t[i_k][p]|$
6: **для** $k = 1, \ldots, |V_i|$
7: **если** $|G_t$[i][p_{best}]-$G_t[i_k][p]| >$ M **то**
8: G_t[i][p_{best}] $\leftarrow$G$_t[i_k][p]$, G_t[i][$best$] $\leftarrow i_k$, G_t[i][$best\_val$] $\leftarrow F(G_t[i_k][p])$
9: **если** Выдать ранжирование **то**
10: Пары вершин (i, G_t[i][$best$]) сортируются по убыванию G_t[i][$best\_val$] и выдаются.
11: **если** Протестировать вершину k **то**
12: **если** Вершина k – больна **то**
13: $G_{t-1}[k][p] \leftarrow 1$, G_t пересчитывается алгоритмом
14: **если** Вершина k – здорова **то**
15: $G_{t-1}[k][p] \leftarrow 0$, G_t пересчитывается алгоритмом

Замечание 4.8. У этого алгоритма есть проблема, выраженная в том, что проверка вершины меняет вероятности не только для рассматриваемой вершины, но и для других вершин, смежных с проверяемой, однако он показывает как можно максимизировать вклад от конкретного слагаемого энтропии системы.

Мы можем построить следующую карту рекомендаций проверки (Рис. 7), где по оси X отложена суммарный риск соседей, по оси Y риск рассматриваемой вершины на прошлой итерации.

Тогда в зависимости от этих двух величин можно понять соседа с какой вероятностью необходимо проверить, чтобы саккумулировать больше энтропии в одной вершине.

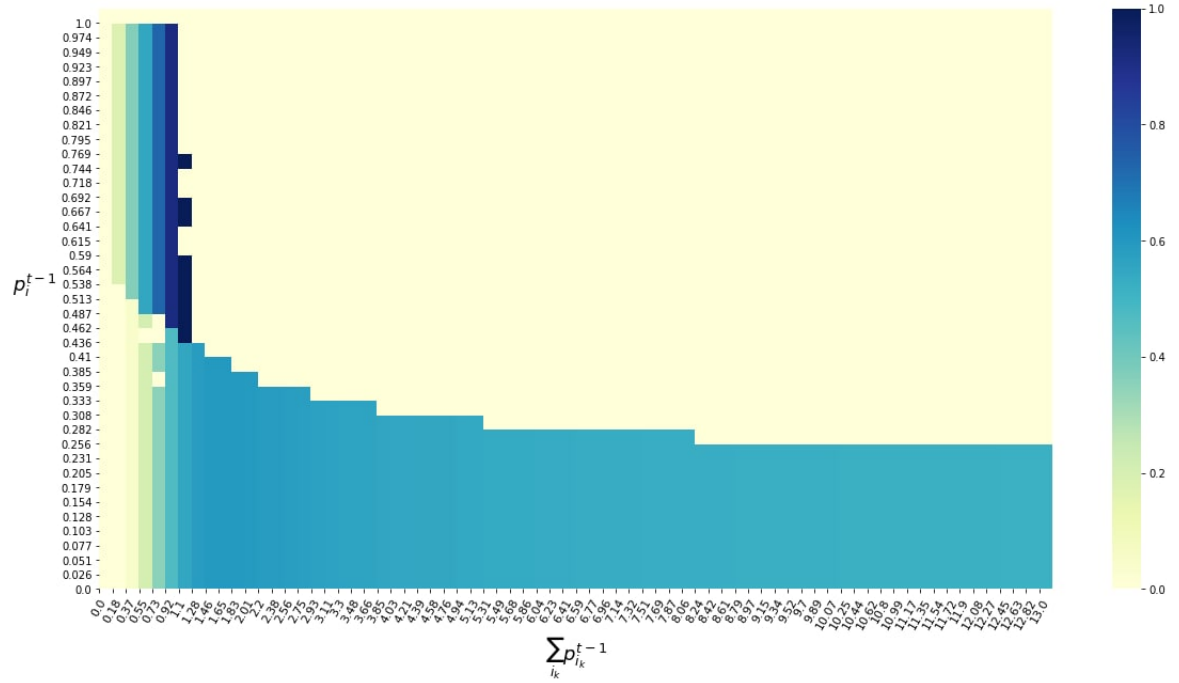

Рис. 7: Карта глубины для вероятности, являющейся оптимальной как для соседа вершины i с точки зрения аккумуляции энтропии.

Замечание 4.9. Такой подход нужен, например, в случае если необходимо отранжировать вершины по принципу важности для наибольшего числа соседей, тем самым увеличив разрыв между слагаемыми энтропии системы, соответствующих разным вершинам.

5 Вычислительные эксперименты

Эксперименты проводились на графах контактов, полученных при помощи системы «Амулет»[20]. Система представляет из себя брелки-маячки, которые соединяются друг с другом при помощи технологии Bluetooth, и сервера, принимающего сигналы от брелков.

Если один маячок попадает в зону досягаемости другого, то между ними формируется связь, отчет о которой они отправляют на сервер. После этого, на основе отчетов, формируется время контакта маячков.

На основе этих данных, для рассматриваемой задачи, были составлены графы контактов всех дней, путём добавления каждого зафиксированного контакта в качестве ребра графа. Если в течение временного периода было несколько контактов одних и тех же маячков, то время контакта складывалось.

Итак, для решения задач был сформирован набор данных, состоящий из 52 графов, соответствующих 52 дням, в течении которых, данная система размещалась в одном из технических предприятий. На рёбрах графов размещалось суммарное время контактов между вершинами в течение соответствующего дня.

5.1 Позитивное и негативное влияние локдауна

В большинстве случаев, введение локдауна – это крайняя мера, которая воспринимается как лучший вариант противостояния эпидемии. Чаще всего это действительно помогает в сдерживании эпидемии и поведение числа заболевших с течением времени выглядят как на показано на (Рис. 8).

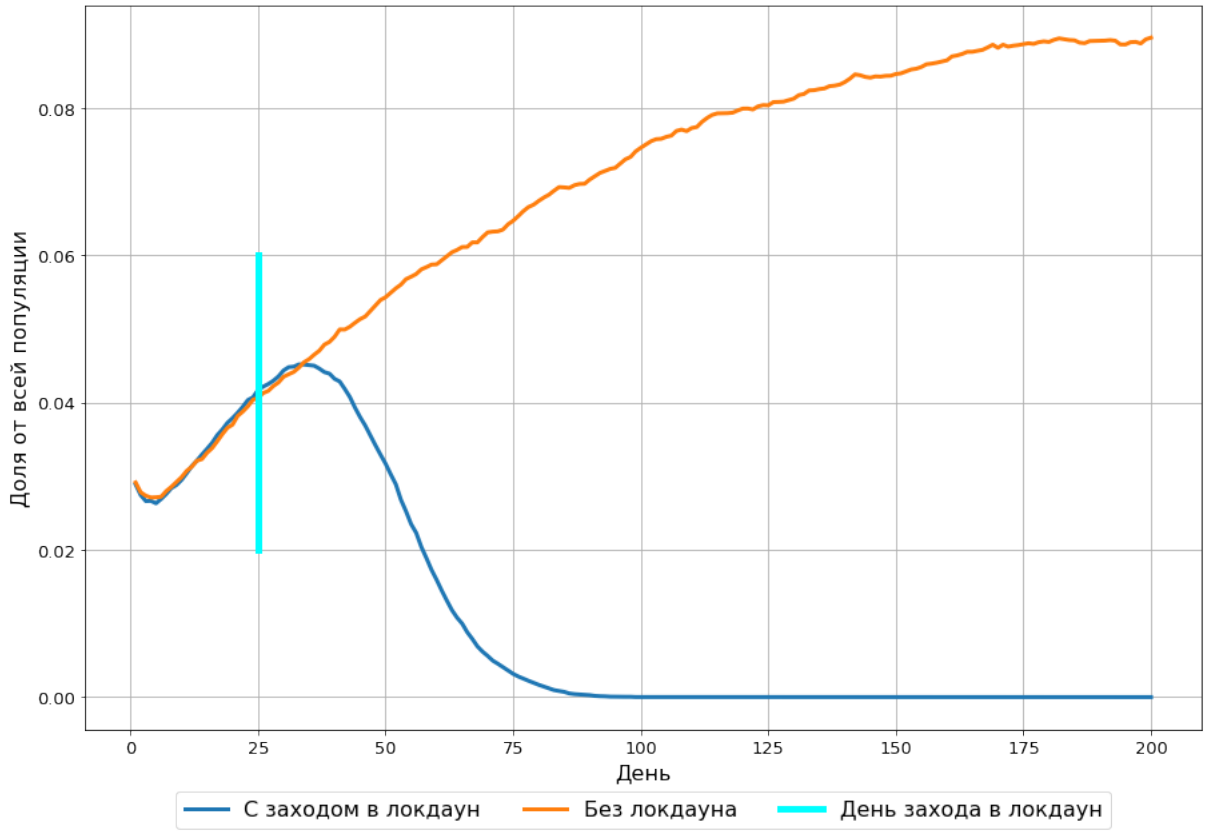

Рис. 8: Поведение эпидемии при положительном влиянии локдауна на ожидаемое число больных.

Однако возможны случаи, когда введение локдауна, напротив, увеличит число заболевших в пике. Этот эффект был назван – эффект локдауна. Эффект, продемонстрированный на (Рис. 9), был отловлен на модельных данных. Данный результат подтверждает теоретические выводы.

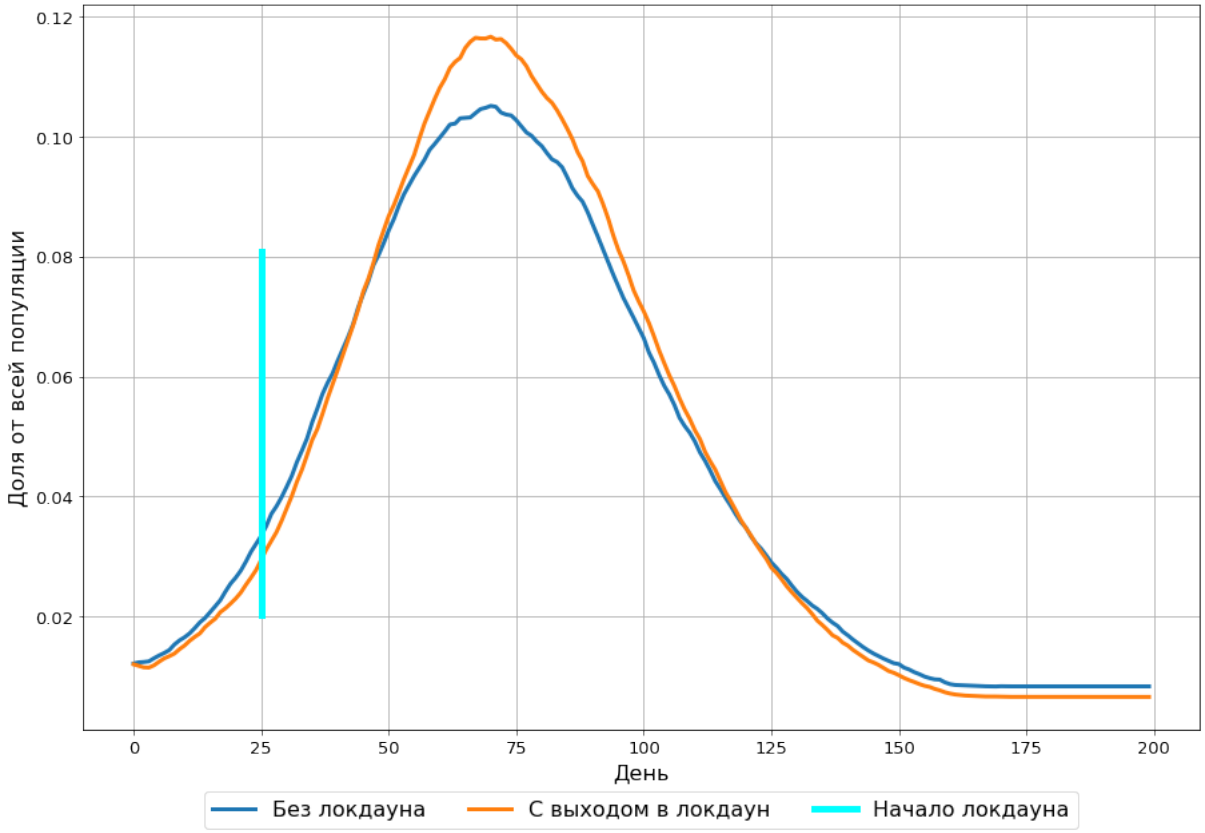

Рис. 9: Поведение эпидемии показывающее существование эффекта локдауна. В эксперименте использовались графы, состоящие из 495 вершин. Когда был период рабочего времени, использовался граф, полученный гауссовским случайным разбиением, а в период домашнего времени, использовался граф, состоящий из клик размера 3. В период рабочего времени β была в $\frac{5}{3}$ раза меньше, чем в период домашнего времени.

5.2 Различные ограничительные меры

Были проведены эксперименты, показывающие монотонное влияние вакцинации и тестирования с изоляцией на распространение эпидемии, а также был представлен не обучаемый метод выбора вершин для вакцинации.

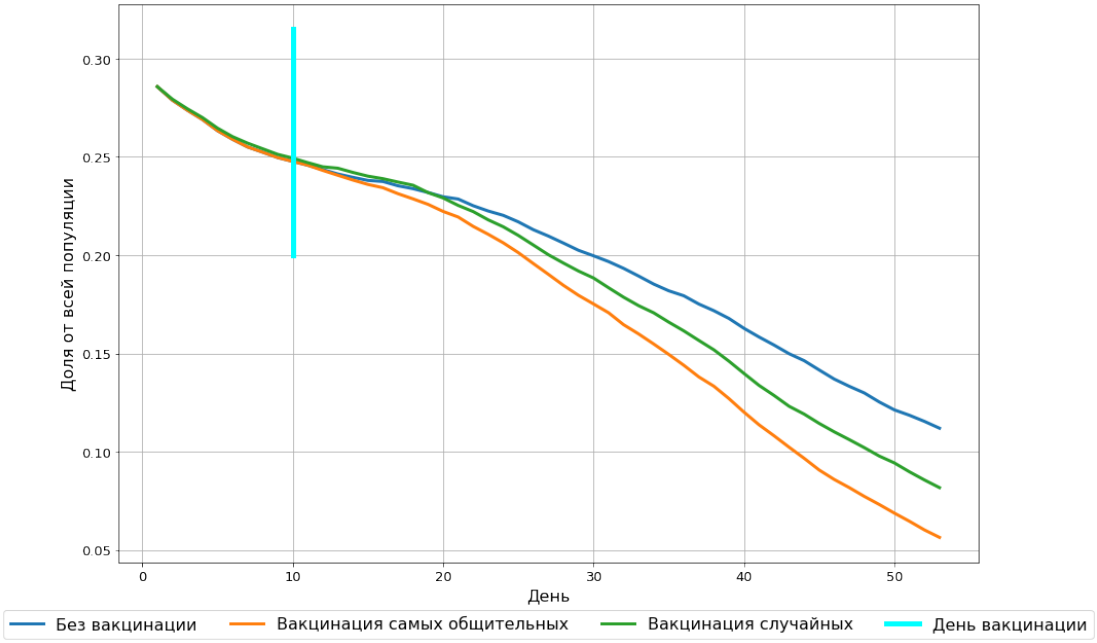

Рис. 10: Число заболевших при различных стратегиях вакцинирования с течением времени. Наибольшее число заболевших в любой момент времени после дня вакцинации у эпидемии, в которой не вакцинировали, чуть меньше заболевших, если вакцинировали случайных, меньше всего заболевших в эпидемии, в которой вакцинировали самых общительные. Самая «общительная» вершина та, которая в прошлом имела больше всего контактов.

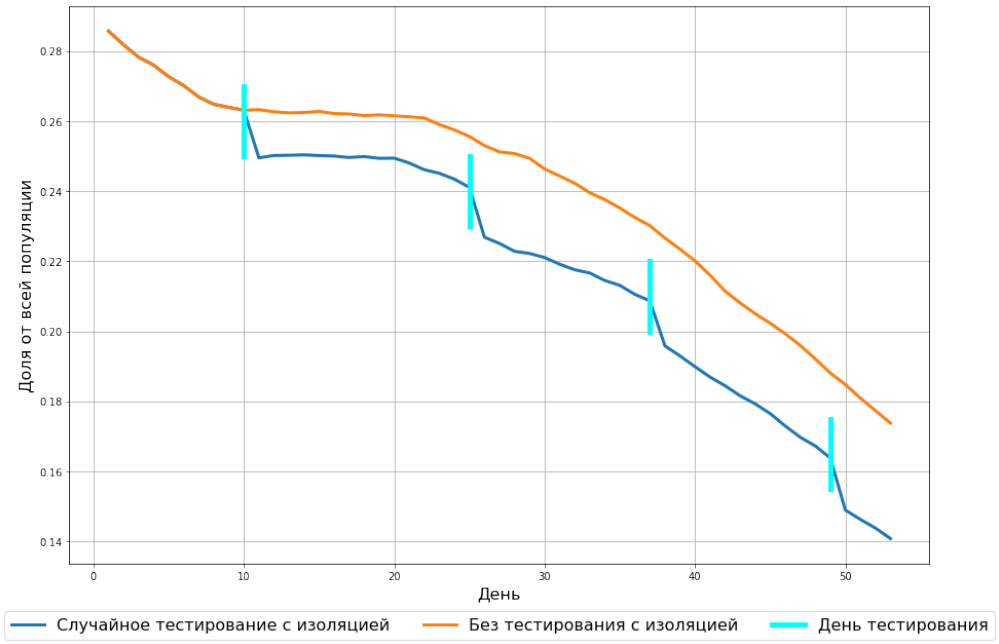

Рис. 11: Число заболевших при двух стратегиях тестирования с изоляции с течением времени. Тестирование случайных и их изоляция в случае если они больны, уменьшает число заболевших на следующих итерациях. Если тестирование показало, что вершина здорова, то ничего не происходит, а если больна, то она заменялась на здоровую.

5.3 Стратегии тестирования

Поскольку алгоритм выбора людей для тестирования (4.1) способен их ранжировать по заданной метрике, было принято решение рассматривать метрики ранжирования NDCG@K, PRECISION@K. Поскольку после тестирования вершины, происходит уточнение вероятностей, также рассматривались метрики классификации, а именно MSE между вероятностями и реальными метками вершин.

Определение 5.1 (Критерий изменения энтропии).

$$\Delta H_i = H - \mathbb{E}_i \tilde{H},$$

где $\mathbb{E}_i \tilde{H}$ *– математическое ожидание изменения энтропии системы, после проверки* i*-ой вершины. Чем больше значение* ΔH_i*, тем выше вершина* i *в рейтинге, то есть выбор согласно этому критерию – это выбор вершины, у которой значение* ΔH_i *максимально.*

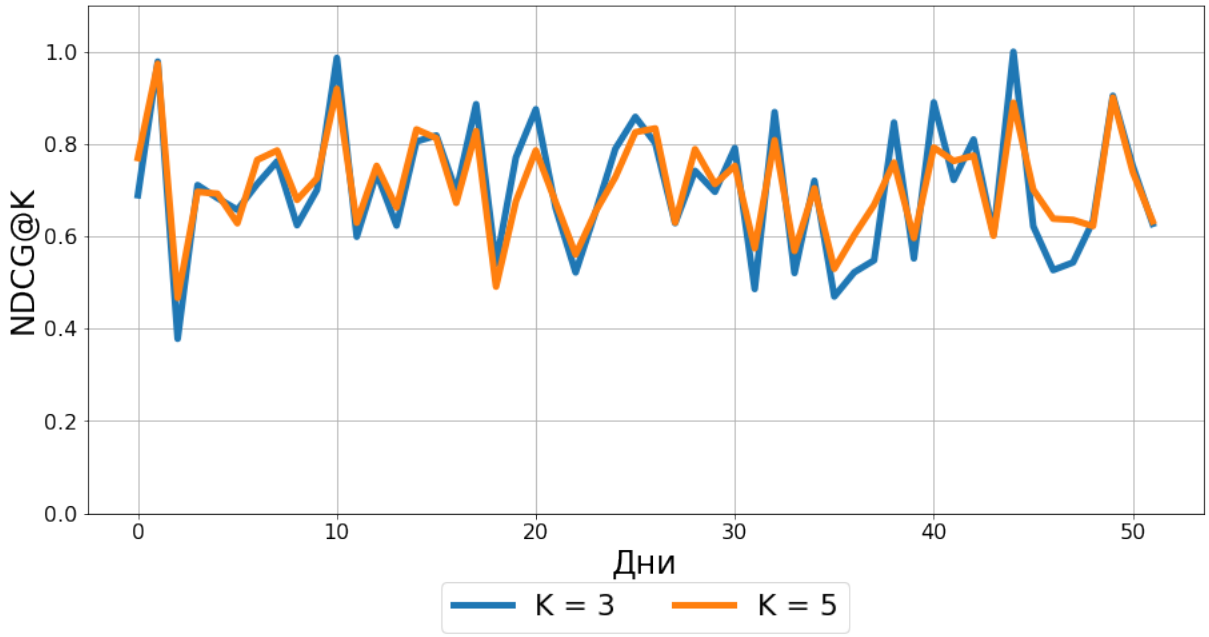

Рис. 12: Среднее значение NDCG на каждом шаге. Ранжирование происходило по критерию изменения энтропии.

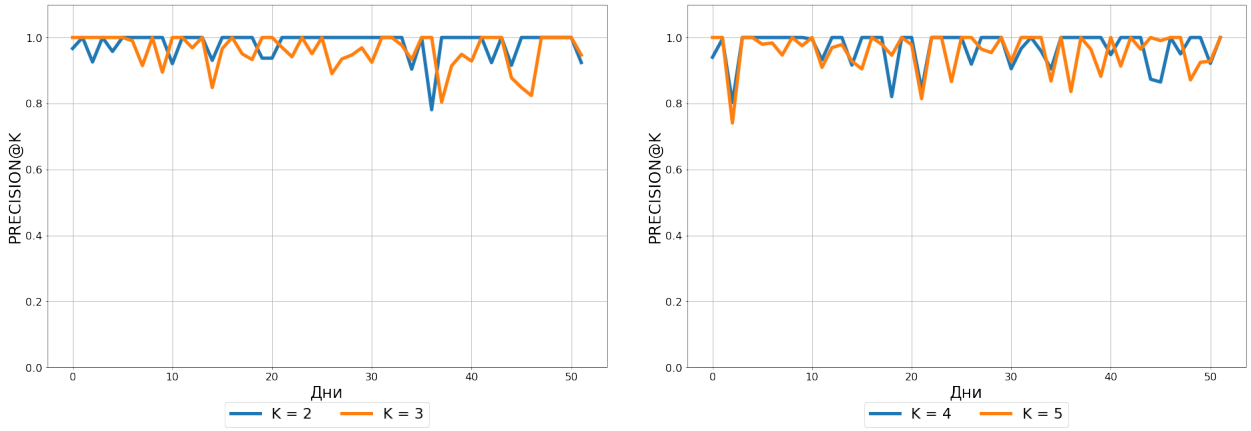

Рис. 13: Средняя значение точности ранжирования вершин по критерию изменения энтропии на каждом шаге.

Несмотря на то, что критерий изменения энтропии оптимален с точки зрения уменьшения энтропии системы, он не является лучшим критерием для приближения вероятностей к меткам. Поэтому были сформулированы критерии, основанные на локальной энтропии и вероятностях меток:

Определение 5.2 (Энтропийный критерий).

$$h_i = H_i + \sum_k H_{i_k},$$

Что есть значение локальной энтропии. Чем больше значение, тем выше вершина i в рейтинге, то есть выбор согласно этому критерию – это выбор вершины, у которой значение h_i максимально.

Определение 5.3 (Вероятностный критерий 1).

$$P_i = \left| p_i - \frac{1}{2} \right| + \sum_k \left| p_{i_k} - \frac{1}{2} \right|$$

Чем меньше значение, тем выше вершина i в рейтинге, то есть выбор согласно этому критерию – это выбор вершины, у которой значение P_i минимально.

Определение 5.4 (Вероятностный критерий 2).

$$\tilde{P}_i = p_i \sum_k \left| \tilde{p}_{i_k} - \frac{1}{2} \right| + (1 - p_i) \sum_k \left| \tilde{\tilde{p}}_{i_k} - \frac{1}{2} \right|,$$

где $\tilde{p}_{i_k}(\tilde{\tilde{p}}_{i_k})$ – вероятности соседей i-ой вершины в случае если i-ая вершина оказалась больна(здорова). Чем меньше значение, тем выше вершина i в рейтинге, то

есть выбор согласно этому критерию – это выбор вершины, у которой значение $\tilde{P}_i$ *минимально.*

Определение 5.5 (Вероятностный критерий 3).

$$\Delta P_i = \tilde{P}_i - P_i,$$

где $\tilde{P}_i$ *– взято из Вероятностного критерия 2, а* P_i *– взято из Вероятностного критерия 1. Чем меньше значение, тем выше вершина* i *в рейтинге, то есть выбор согласно этому критерию – это выбор вершины, у которой значение* ΔP_i *минимально.*

Для того, чтобы протестировать какой из критериев выбора является наилучшим для приближения вероятностей, были проведены эксперименты на 10 реализованных начальных инициализаций больных и в общей сложности 10000 реализациях эпидемий на графе, полученный при помощи системы «Амулет». Результат представлен на (Рис. 14).

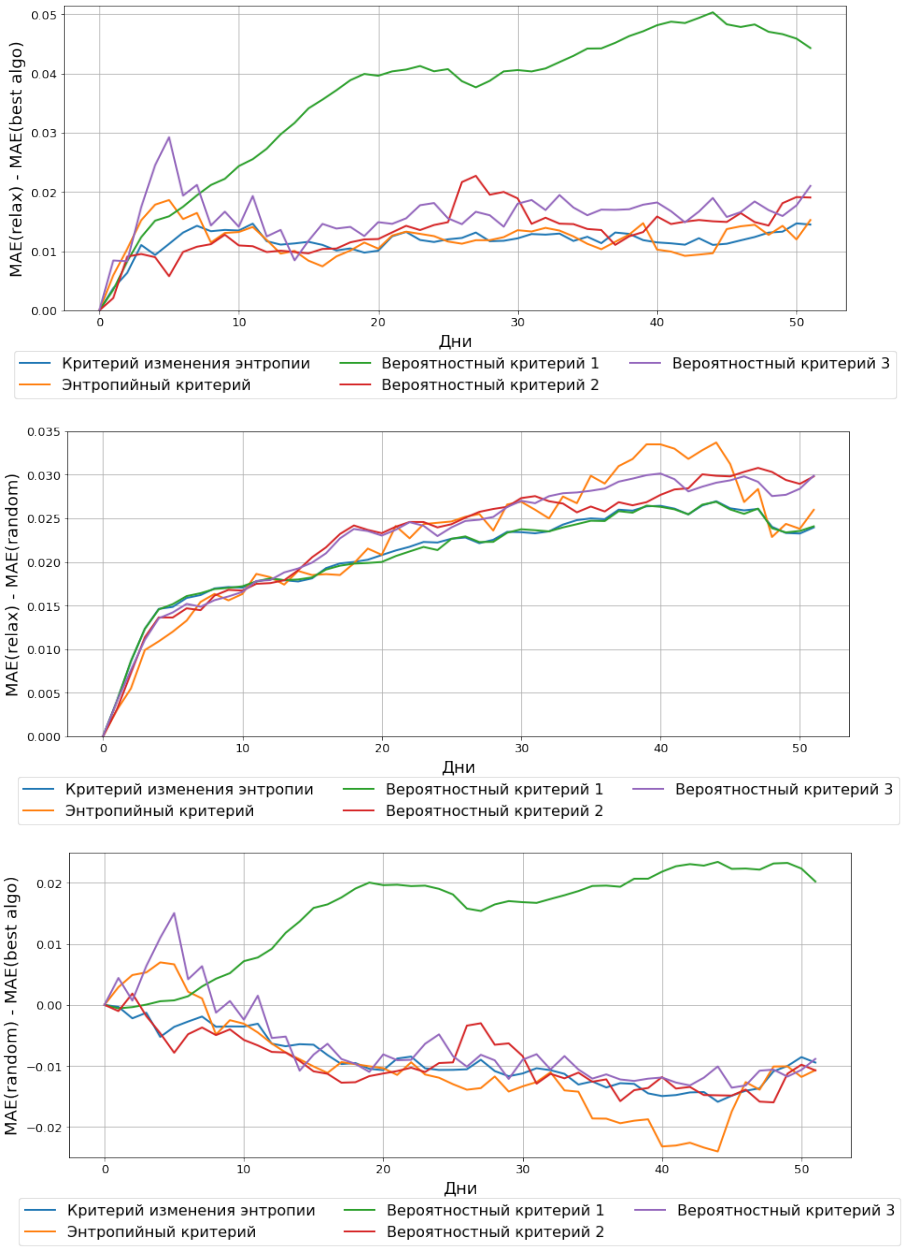

Рис. 14: relax – алгоритм не тестирует вершины, random – алгоритм тестирует случайного человека на каждом шаге, best algo – алгоритм тестирует лучшего человека с точки зрения одного из критериев. MAE(·) – суммарная абсолютная ошибка предсказания меток для всех вершин в конкретный день. Чем больше значение по оси y, тем лучше второй алгоритм (в разности оси) приближает вероятности к меткам по сравнению с первым.

Как видно, выбор лучшего человека согласно Вероятностному критерию 1, наилучшим образом приближает вероятности к реальным меткам вершин.

6 Заключение

В ходе данной работы были рассмотрены вопросы моделирования и анализ распространения эпидемии на графах в случае наличия ограничительных мер и их отсутствия.

Получены следующие результаты:

- Сформулирована и доказана теорема об эквивалентности людей для симуляции эпидемии.
- Предложена модель заражения, учитывающая время и другие свойства контактов для расчета вероятности заражения.
- Предложены модели различных противоэпидемиологических мер: вакцинации, изоляции больных, локдауна.
- Доказано, что вакцинация и изоляция больных не увеличивает математическое ожидание заболевших.
- Показано, что локдаун может как уменьшить, так и увеличить математическое ожидание числа заболевших. Сформулирован и доказан критерий увеличения математического ожидания числа заболевших при локдауне.
- Поставлена серия экспериментов, иллюстрирующих полученные аналитически сценарии развития эпидемии.
- Предложена формулировка задачи тестирования в терминах теории информации.
- Предложены, реализованы и экспериментально проверены алгоритмы тестирования, основанные на предложенных информационных критериях.

В будущем планируется перенести алгоритмы и результаты на смежные области, применить алгоритм аккумуляции энтропии для подзадач тестирования, а также добавить обратное распространение ошибки к текущей постановки задачи.

Список литературы

[1] Burr Settles. «Active learning literature survey». в: (2009).

[2] R. Ross. «An application of the theory of probabilities to the study of a priori pathometry. Part I.» в: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 92(638) (1916), с. 204—230.

[3] T. Harko, Francisco S. N. Lobo и M. Mak. «Exact analytical solutions of the Susceptible-Infected-Recovered (SIR) epidemic model and of the SIR model with equal death and birth rates». в: *Appl. Math. Comput.* 236 (2014), с. 184—194.

[4] Linda J.S. Allen. «Some discrete-time SI, SIR, and SIS epidemic models». в: *Mathematical Biosciences* 124.1 (1994), с. 83—105. ISSN: 0025-5564. DOI: https://doi.org/10.1016/0025-5564(94)90025-6. URL: https://www.sciencedirect.com/science/article/pii/0025556494900256.

[5] Vincenzo Capasso. *Mathematical Structures of Epidemic Systems*. Springer-Verlag Berlin Heidelberg, 1993.

[6] Y. Moreno, R. Pastor-Satorras и Alessandro Vespignani. «Epidemic outbreaks in complex heterogeneous networks». в: *The European Physical Journal B - Condensed Matter and Complex Systems* 26 (2002), с. 521—529.

[7] R. Pastor-Satorras и др. «Epidemic processes in complex networks». в: *ArXiv* abs/1408.2701 (2014).

[8] S. Gómez и др. «Discrete-time Markov chain approach to contact-based disease spreading in complex networks». в: *EPL (Europhysics Letters)* 89.3 (февр. 2010), с. 38009. ISSN: 1286-4854. DOI: 10.1209/0295-5075/89/38009. URL: http://dx.doi.org/10.1209/0295-5075/89/38009.

[9] Bo Qu и Huijuan Wang. *The Accuracy of Mean-Field Approximation for Susceptible-Infected-Susceptible Epidemic Spreading*. 2016. arXiv: 1609.01105 [physics.soc-ph].

[10] Lauren Ancel Meyers. «Contact network epidemiology: Bond percolation applied to infectious disease prediction and control». в: *Bull. Amer. Math. Soc.* 44 (окт. 2006), с. 63—86. URL: https://doi.org/10.1090/S0273-0979-06-01148-7.

[11] Jack Leitch, Kathleen A Alexander и Srijan Sengupta. «Toward epidemic thresholds on temporal networks: a review and open questions». в: *Applied Network Science* 4.1 (2019), с. 1—21.

[12] Abby Leung и др. *Contact Graph Epidemic Modelling of COVID-19 for Transmission and Intervention Strategies*. 2020. arXiv: `2010.03081 [cs.SI]`.

[13] Manlio De Domenico и др. «The anatomy of a scientific rumor». в: *Scientific reports* 3.1 (2013), с. 1—9.

[14] Aron Culotta и Andrew McCallum. «Reducing labeling effort for structured prediction tasks». в: *AAAI*. т. 5. 2005, с. 746—751.

[15] Burr Settles и Mark Craven. «An analysis of active learning strategies for sequence labeling tasks». в: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. 2008, с. 1070—1079.

[16] Tobias Scheffer, Christian Decomain и Stefan Wrobel. «Active hidden markov models for information extraction». в: *International Symposium on Intelligent Data Analysis*. Springer. 2001, с. 309—318.

[17] Claude E Shannon. «A mathematical theory of communication». в: *The Bell system technical journal* 27.3 (1948), с. 379—423.

[18] Rayan SM. *SEIRSplus*. `https://github.com/ryansmcgee/seirsplus`. 2020.

[19] Stephen Boyd, Stephen P Boyd и Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[20] softtree. *Amuleit - Collective immunity management and monitoring*. 2020. URL: `https://amuleit.ru` (дата обр. 14.06.2021).