ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет информатики, математики и компьютерных наук

Программа подготовки бакалавров по направлению

01.03.02 Прикладная математика и информатика

Артамонов Денис Сергеевич

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Исследование статистической неопределённости знаковых процедур для построения сетевых структур и характеристик фондового рынка

Рецензент

д.ф.-м.н., профессор

Колданов А.П.

Научный руководитель

Преподаватель

Д.П. Семёнов

Нижний Новгород, 2020

Содержание

1 Вступление **2**

2 Сетевой анализ рынка **5**

2.1 Процедуры фильтрации 8

2.2 Пример 11

3 Меры неопределённости **15**

3.1 Мера доли ошибок($\mathcal{E}$-мера) 16

4 Эксперименты **18**

4.1 Анализ эталонной сети 24

4.2 Измерение неопределённости сетевых структур 27

4.2.1 Максимальное остовное дерево 28

4.2.2 Рыночный граф 31

4.2.3 Максимальная клика 34

4.2.4 Максимальное независимое множество 37

5 Заключение **40**

1 Вступление

Сегодня сетевой анализ фондовых рынков является очень активной областью исследования [14] [10] [11] . Фондовый рынок можно представить как полный ненаправленный взвешанный граф, в котором вершины соответствуют исследуемым акциям, а веса рёбер рассчитваются на основе некоторой меры близости, например корреляции Пирсона доходностей соответствующих акций. Таким образом, имея подобное представление в виде гарфа, мы можем использовать всю мощь теории графов для анализа и исследования закономерностей и трендов фондовых рынков, а также для составления более доходных портфелей. Так, в работе [13] сетевой анализ был применён для анализа российского рынка ценных бумаг, а в работе [5] - для китайского.

Однако, полный граф представляет собой достаточно сложную структуру для анализа. Например, полный граф со 100 вершинами имеет 4950 рёбер, а граф с 500 вершинами уже 124750 рёбер. Для упрощения такой структуры, используются различные процедуры фильтрации, позволяющие преобразовать исходный граф в иную сетевую структуру с меньшим числом рёбер, которая содержит лишь наиболее значимую для анализа информацию. Такими процедурами могут быть: построение максимального остовного дерева(maximum spanning tree)[8], рыночного графа(market graph)[2], а также клики и независимого множества максимального размера (maximum clique

and maximum independent set), построенные на основе рыночного графа. В результате применения каждой их этих процедур, мы получаем подграф исходного рыночного графа, содержащий меньшее число рёбер. При этом оставшиеся отражают наиболее сильные, относительно принятой меры близости, зависимости между акциями.

Работа с сетевыми структурами при анализе фондового рынка должна включать оценку статистической неопределённости из-за стохастической природа временных рядов, котрыми являются данные о доходности акций. Сетевые структуры, построенные на основе данных из различных временных промежутков, могут различаться, а значит анализ подобных структур может приводить к разным выводам в зависимости от измерений. Таким образом, актуальным становится вопрос: насколько можно доверять выводам, сделанным на основе анализа той или иной структуры?

В работе [7] была изучена статистическая неорпеделённость различных процедур фильтрации. В ней были предложены две меры для оценки статистической неопределённости сетевых структур. Относительно одной из предложенных мер, экспериментальным путём была рассчитана статистическая неопределённость различных процедур фильтрации, а также была найдена наиболее стабильная относительно статистичекой неопределённости сетевая структура. Ею оказалась максимальная клика.

В работе [7] предполагалось, что доходности имели нормальное

распределение, а мерой близости для создания сети стала выборочная корреляция Пирсона, и статистическая неопределённость была рассчитана только для таких начальных условий.

Цель моей работы - расширить результаты полученные в [7], проведя серию экспериментов для оценки статистической неопределённости процедур фильтрации при иных начальных предположениях, а именно:

1. Использование корреляции Пирсона в качестве меры близости, предполагая, что доходности акций имеют смешанное распределение нормального и распределения Стьюдента

2. Использование вероятности совпадения знаков в качестве меры близости, предполагая, что доходности акций имеют смешанное распределение нормального и распределения Стьюдента

Предположение о смешанном распределении позволит нам рассмотреть поведение неопределённости структур при постепенном отклонении от нормального распределения. Использование же вероятности совпадения знаков в качестве меры близости даст возможность изучить статистическую неопределённости структур иного рода сети, основанной на иной мере близости.

Структура работы следующая. В разделе 2 мы формально определим сетевую структуру фондового рынка, детально опишем меры близости, а также процедуры фильтрации и алгоритмы для их при-

менения. Также, будет приведён пример построения сети и применения к ней различных процедур фильтрации, основанный на реальных данных. В разделе 3 мы перейдём к рассмотрению подходов к изучению статистической неопределённости: изучим общий подход к оценке неопределённости, основанный на сравнении эталонной и выборочной сетей, определим меры неопределённости, формализуем понятия эталонной и выборочной сетей, а также понятие стабильности сетевой структуры относительно неопределённости. В разделе 4 представлены результаты экспериментов, а также их детальное описание. В последнем разделе подведём итоги и сделаем вывод и полученных результатах.

2 Сетевой анализ рынка

Прежде чем приступить к изучению мер неопределённости, рассмотрим вопрос построения рыночных сетевых структур на основе разных мер близости. Предположим, что мы хотим построить граф, основываясь на наблюдениях за N акциями в течении n дней. Данные представляют собой цены акций в момент закрытия торгов в каждый из дней. В первую очередь, имеющиеся данные мы преобразуем в данные о ежедневной доходности акций. Доходность акции k в день t рассчитывается как

$$R_k(t) = \ln \frac{P_k(t)}{P_k(t-1)} \tag{1}$$

где $P_k(t)$ - цена акции k в день t.

Предположим, что для некоторого фиксированного k и $t = 1, ..., nR_k(t)$ - независимые одинаково распределённые случайные величины и имеют такое же распределение, что и случайная величина $R_k, k = 1..N$. Можно также предположить, что случайные величины $R_1, ..., R_N$ имеют совместное нормальное распределение. В этом случае выборочная матрицы корреляции $||r_{ij}||$ будет содераржать большую часть доступной информации о зависимости этих случайных величин[12]. Каждый элемент такой матрицы вычисляется как

$$r_{ij} = \frac{\sum_{t=1}^{n}(R_i(t) - \overline{R_i})(R_j(t) - \overline{R_j})}{\sqrt{\sum_{t=1}^{n}(R_i(t) - \overline{R_i})^2}\sqrt{\sum_{t=1}^{n}(R_j(t) - \overline{R_j})^2}} \tag{2}$$

где $\overline{R_i} = \frac{1}{n}\sum_{t=1}^{n} R_i(t)$ - среднее R_i. Такая матрица часто используется в качестве меры близости в сетевой структуре, отражающей структуру фондового рынка.

Однако, если появляются какие-то отклонения совместного распределения случайных величин от нормального, матрица $||r_{ij}||$ перестанет полностью характеризовать взаимосвязь между случайными величинами. Изучение этих отклонений и влияние их на статистическую неопределённость сетевых структур являются одной из целей данной работы.

Чтобы смоделировать подобные отклонения, предположим, что доходности имеют смешанное распределение: часть наблюдений рас-

пределена согласно нормальному закону, а остальные имеют распределение Стьюдента. Введём параметр $r \in [0, 1]$, который будет отвечать за то, какая доля наблюдений иметь распределение Стьюдента.

Один из общих подходов построения многомерного распределения Стьюдента (или t-распределения) размерности p, описанный в [3], состоит в следующем. Пусть Y и u - независимые и распределеные как $Y \sim N(0, \Sigma)$ и $u \sim \chi^2_\nu$, где Σ - квадратная матрица размерностью $p \times p$. Пусть T такая что

$$\frac{Y}{\sqrt{u/\nu}} = T - \mu \tag{3}$$

Тогда T иммет многомерное распределение Стьюдента с параметрами Σ, μ, ν: $T \sim t_\nu(\mu, \Sigma)$. Заметим, что матрица Σ не является матрицей ковариации. Матрица ковариации определяется как $\frac{\nu}{\nu-2}\Sigma$ для $\nu > 2$. В этой работе будем использовать параметр $\nu = 3$.

В работе [1] была представлена мера близости, которую можно использовать в условиях подобных отклонений. Эта мера основана на вероятности совпадений знаков доходностей. Она позволяет описать совместное поведение случайных величин $R_1, ..., R_N$ даже если совместное распределение случаных величин неизвестно. Она определяется как

$$p_{ij} = P(R_i \geq E(R_i), R_j \geq E(R_j) \cup R_i < E(R_i), R_j < E(R_j)) \tag{4}$$

В этой работе мы будем использовать схожую меру - выборочная

вероятность совпадения :

$$\gamma_{ij} = \frac{1}{n} \sum_{t=1}^{n} I[R_i(t) R_j(t)] \tag{5}$$

где

$$I[x] = \begin{cases} 1, & \text{если } y \geq 0 \\ 0, & \text{если } y < 0 \end{cases}$$

Фондовый рынок можно представить как конечный полный взвешенный неориентированный граф без петель и кратных рёбер $G = (V, E)$, где $V = 1, 2, ..., N$ - множество вершин, соответствующее наименованиям акций, а s_{ij} - вес ребра $(i, j) \in E$, отражающий значение некоторой меры близости соответствующих акций. Такой граф мы будем называть рыночной сетевой структурой. В этой работе будут рассмотрены структуры, основанные на двух мерах близости: выборочной корреляции Пирсона и выборочной вероятности совпадения знаков. В зависимости от выбранной структуры, матрицей весов графа будет либо $||r_{ij}||$ для выборочной корреляции Пирсона, либо $||\gamma_{ij}||$ для выборочной вероятности совпадения знаков.

2.1 Процедуры фильтрации

Для извлечения наиболее ценной информации из сетевой структуры и упрощения дальнейшего анализа используются различные процедуры фильтрации, которые предполагают построение подграфа исходной сети. Мы будем рассматривать следующие процедуры:

1. Построение максимального остовного дерева (maximum spanning tree)

2. Построение рыночного графа (market graph)

3. Построение максимальной клики на основе рыночного графа (maximum clique)

4. Построение максимального независимого множества на основе рыночного графа (maximum independent set)

Оставное дерево связанного графа представляет собой связанный подграф, содержащий все вершины исходного подграфа и не имеющий циклов. Максимальное остовное дерево - это оставное дерево, имеющий максимальный вес, который вычисляется как сумма весов ребёр, входящих в остовное дерево. Мы рассматриваем именно максимальное остовное дерево, так как заинтересованы в том, чтобы найти наиболее связанные между собой акции относительно показателей их доходностей. Использование этой процедуры было предложено в [9].

Позднее, в работе [2] был представлен и изучен рыночный граф (market graph), который получается из исходного путём удаления всех рёбер, веса которых меньше некоторого заданного значения порога θ. В полученном графе можно найти максимальную клику и максимальное независимое множество, с помощью которых можно

проанализировать структуру фондового рынка, а максимальное независимое множество можно рассматривать как основу для создания диверсифицированного портфеля акций. В неориентированном графе кликой называют подмножество вершин, которые порождают полный подграф исходного графа. Максимальная клика графа - это клика максимального размера, то есть включающая в себя максимально возможное число вершин. Независимое множество - понятие, обратное понятию клики. Это подмножество вершин графа, не имеющих общих рёбер. Максимальное независимое множество - это независимое множество максимального размера.

При построении рыночного графа, а также максимальной клики и максимального независимого множества на основе сети, веса в которой представляют собой выборочную вероятность совпадения знаков возможно установить связь между классической корреляцией Пирсона и корреляцией знаков[1]. Обе меры связаны соотношением

$$\theta_\gamma = \frac{1}{2}\left(\frac{2}{\pi}\arcsin\theta_r + 1\right) \tag{6}$$

Таким образом, мы сможем сравнить статистическую неопределённость структур, построенных на основе разных мер близости. Выбрав порог θ_r для построения рыночного графа на основе корреляции Пирсона, мы можем расчитать порог θ_γ для рыночного графа на основе корреляции знаков. Меры неопределённости для полученных структур будут сравнимы дург с другом.

2.2 Пример

Рассмотрим пример построения сети на основе корреляции Пирсона для следующего набора акций NASDAQ:

1. AAPL (Apple)
2. AMZN (Amazon)
3. CZR (Caesars Entertainment Corporation)
4. GOOG (Alphabet Inc.)
5. LBTYA (Liberty Global)
6. NAVI (Navient)
7. PEP (PepsiCo Inc.)
8. PTLA (Portola Pharmaceuticals)
9. SFM (Source Filmmaker)
10. XOG (Extraction Oil & Gas)

В качестве меры близости будем использоват корреляцию Пирсона. Имея данные о ценах этих акций за 2018-2019, рассчитаем их доходности и составим матрицу выборочной корреляции Пирсона $||s_{ij}||$. Матрица будет иметь вид:

	AAPL	*AMZN*	*CZR*	*GOOG*	*LBTYA*	*NAVI*	*PEP*	*PTLA*	*SFM*	*XOG*
AAPL	1.000	0.626	0.337	0.631	0.309	0.354	0.270	0.352	0.151	0.326
AMZN	0.626	1.000	0.451	0.702	0.334	0.337	0.181	0.341	0.152	0.279
CZR	0.337	0.451	1.000	0.422	0.278	0.358	0.143	0.276	0.086	0.232
GOOG	0.631	0.702	0.422	1.000	0.344	0.327	0.244	0.384	0.157	0.242
LBTYA	0.309	0.334	0.278	0.344	1.000	0.338	0.175	0.206	0.029	0.258
NAVI	0.354	0.337	0.358	0.327	0.338	1.000	0.156	0.276	0.189	0.310
PEP	0.270	0.181	0.143	0.244	0.175	0.156	1.000	0.115	0.042	0.075
PTLA	0.352	0.341	0.276	0.384	0.206	0.276	0.115	1.000	0.077	0.201
SFM	0.151	0.152	0.086	0.157	0.029	0.189	0.042	0.077	1.000	0.105
XOG	0.326	0.279	0.232	0.242	0.258	0.310	0.075	0.201	0.105	1.000

Тепловая карта матрицы представлена на рисунке 1.

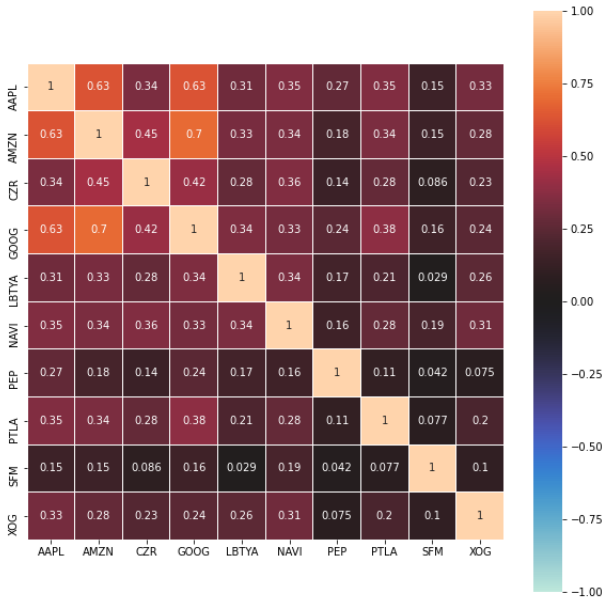


Рис. 1: Тепловая карты матрицы корреляций

Соответствующий матрице граф представлен на рисунке 2.

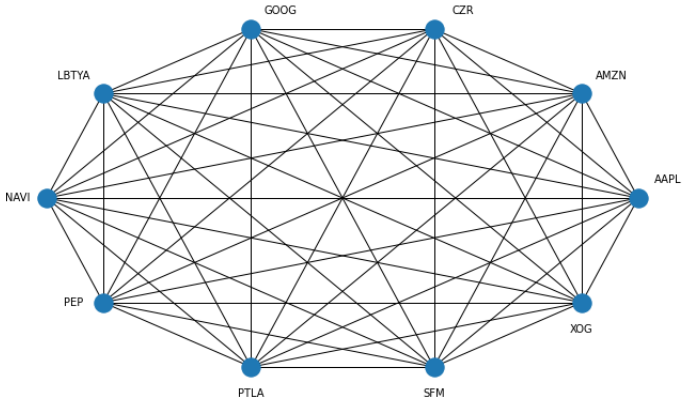


Рис. 2: Сетевая структура

Применим различные процедуры фильтрации.Из полученной сети построим максимальное остовное дерево.

Рис. 3: Максимальное остовное дерево

Построим рыночный граф с порогом $\theta = 0.3$.

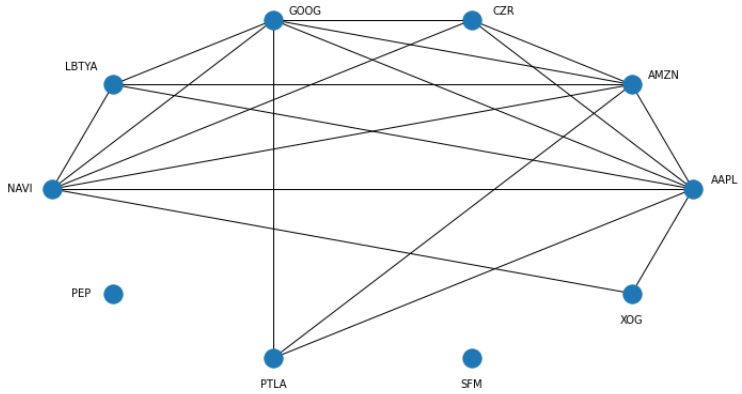


Рис. 4: Рыночный граф

В полученном рыночном графе найдём максимальную клику и максимальное независимое множество.

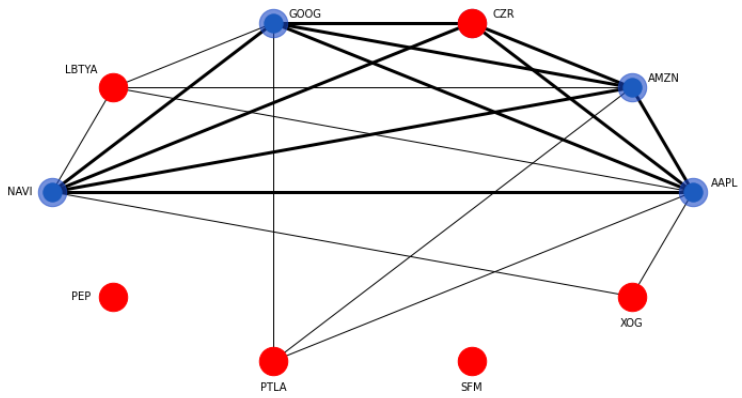


Рис. 5: Максимальные клика и независимое множество

Максимальная клика содержит 5 вершин, соединённых на рисунке 5 жирными линиями. Максимальное независимое множество содержит 6 вершин, одна из которых присутствует и в клике. Вер-

шины, входящие в максимальное независимое множество отмечены на рисунке 5 красным цветом.

3 Меры неопределённости

Построив сетевую структуры на одном наборе наблюдений, мы хотим быть уверены, что она будет актуальна и для другого набора. Для этого мы измеряем статистическую неопределённость структуры. И чем меньше эта неопределённость, тем с большей мы будем уверены, что она сохранит свой вид, а значит выводы, полученные на её основе будут более точными.

Предполагается, что есть некотрая эталонная сетевая структура, построенная на основе параметром распределения. Например, предполагая, что случайные величины $R_1, .., R_N$ имеют совместное нормальное распределение, мы можем сказать, что их взаимоотношения описываются матрицей корреляций Пирсона $||\rho_{ij}||$. Тогда сетевая структура, полученная на основе этой матрицы будет эталонной [7]. Так как мы не знаем точно, какова матрица $||\rho_{ij}||$, мы можем оценить её матрицей выборочной корреляции $||r_{ij}||$. Тогда сетевую структуру, полученную на основе матрицы $||r_{ij}||$ мы также будем называть эталонной и предполагать, что она отражает истинные взаимоотношения между $R_1, .., R_N$, а любая другая сетевая структура, может приблизить эталонную структуру с той или иной точностью.

В работе [7] был предложен единый подход к измерению неопределённости сетевых структур. Он заключается в сравнении эталонной структуры, которая строится на основе имеющихся данных, и некоторой выборочной структуры, построенной на основе данных, сгенерированных из некоторого распределения. Таким образом, мы можем изучить, насколько "сложно"сетевой структуре приблизить эталонную структуру с некоторой точностью, а именно, сколько наблюдений понадобится, чтобы получить истинную структуру с некоторым заданным порогом ошибки. Также в [7] были предложены $\mathcal{R}$ и $\mathcal{E}$ меры близости, и на основании $\mathcal{E}$-меры были проведены эксперименты по оценке статистической неопределённости структур. В нашей работе также будет использована $\mathcal{E}$-мера.

3.1 Мера доли ошибок($\mathcal{E}$-мера)

$\mathcal{E}$-меру также называют мерой доли ошибок. Она основана на рассчёте количества рёбер, которые ошибочно входят или в выборочную структуру или отсутствуют в ней в сравнении с эталонной сетью. Определяется эта мера следующим образом.

Пусть

$$x_1^{ij} = \begin{cases} 1, & \text{если } (i,j) \text{ ребро ошибочно включено в выборочную структуру,} \\ 0, & \text{в противном случае} \end{cases}$$

и

$$x_2^{ij} = \begin{cases} 1, & \text{если } (i,j) \text{ ребро ошибочно отсутствует в выборочной структуре,} \\ 0, & \text{в противном случае} \end{cases}$$

Также определим

$$X_1 = \sum_{1 \le i \le j \le N} x_1^{ij}, \quad X_2 = \sum_{1 \le i \le j \le N} x_2^{ij}$$

Таким образом, получается, что X_1 - число рёбер, которые содержит выборочная структура, но не содержит эталонная, а X_2 - наоборот, число рёбер, которые содержит эталонная структура, но не содержит выборочная.

Определим случайную величину X, такую что

$$X = \frac{1}{2}\left(\frac{X_1}{M_1} + \frac{X_2}{M_2}\right), \tag{7}$$

где M_1 - максимально возможное значение X_1, то есть число рёбер в выборочной структуре, а M_2 - максимально возможное значение X_2, то есть число рёбер в эталонной структуре. Случайная величина X принимает значение $X \in [0,1]$ и описывает общую долю ошибок. Тогда $\mathcal{E}(\mathcal{S}, n)$-мерой статистической неопределённости для некоторой сетевой структуры $\mathcal{S}$ будет являться математическое ожидание этой случайной величины:

$$\mathcal{E}(\mathcal{S}, n) = E[X] \tag{8}$$

Если значение $X_1 = 1$, все рёбра в выборочной структуре включены неверно, если $X_2 = 1$, ни одно ребро из эталонной структуры

не входит в выборочную. $\mathcal{E}$-мера называют $\mathcal{E}$-мерой уровня $\mathcal{E}_0$, если при числе наблюдений $n_\mathcal{E}$: $\mathcal{E}(\mathcal{S}, n_\mathcal{E}) = \mathcal{E}_0$.

Говорят, что структура $\mathcal{S}_1$ стабильнее, чем структура $\mathcal{S}_2$, если $\mathcal{E}(\mathcal{S}_1, n) < \mathcal{E}(\mathcal{S}_2, n)$ для любого числа наблюдений n. Иными словами, статистическая непределённость структуры $\mathcal{S}_1$ меньше, чем статистическая непределённость структуры $\mathcal{S}_2$, если $\mathcal{E}(\mathcal{S}_1, n_1) = \mathcal{E}(\mathcal{S}_2, n_2)$, когда $n1 = n2$.

4 Эксперименты

В этой части работы мы будем экспериментало изучать статистическую неопределённость различных сетевых структур. Перед нами стоят две задачи: понять, как отклонение от нормального распределения влияет на статистическую неопределённость и изучить статистическую неопределённость структур, построенных на основе новой меры близости - вероятности совпадения знаков.

Таким образом, экспериментальная часть будем состоять из двух частей. В первой части в качестве меры близости мы применяем выборочную корреляцию Пирсона, во второй - вероятность совпадения знаков. Каждая из частей эксперимента включает в себя оценку статистической неопределённости каждой из структур при различных значениях коэффициента r смешанного распределения. Список изучаемых структур придставлен ниже:

- Максимальное остовное дерево (maximum spanning tree)
- Рыночный граф (market graph)
- Максимальная клика на основе рыночного графа (maximum clique)
- Максимальное независимое множество на основе рыночного графа (maximum independent set)

Как и в [7] мы будем оценить статистическую неопределённость, основываясь на $\mathcal{E}(\mathcal{S}, n)$-мере по следующему алгоритму:

1. Построить эталонную структуру на основе имеющихся данных
2. Сгенерировать выборку доходностей акций $x_{11}, ..., x_{1N}, ..., x_{n1}, ..., x_{nN}$
3. Рассчитать матрицу весов на основе меры близости и сгенерированных наблюдений
4. Построить выборочную сеть
5. В выборочной сети построить выборочную сетевую структуру
6. Рассчитать долю ошибок I($\frac{X_1}{M_1}$) и II рода($\frac{X_2}{M_2}$) , а также общую долю ошибок X
7. Повторить 500 раз шаги 2-5 и рассчитать среднее общей доли ошибок X. Полученное значение и будет являться оценкой меры $\mathcal{E}(\mathcal{S}, n)$.

Генерация выборки доходностей акций из смешанного распределения происходит следющим образом. В функцию генерации `mixed_t_normal()` кроме параметров, необходимых для генераторов из многомерного нормального распределения и многомерного распределения Стьюдента, передаётся число наблюдений n и параметр $r \in [0, 1]$, который отвечает за то, какая часть из n наблюдений будет сгенерирована из многомерного распределения Стьюдента. Оставшая часть будет сгенерирована из многомерного нормального распределения. Например, если при $n = 100$ и $r = 0.3$ 30 наблюдений будет сгенерировано из многомерного распределения Стьюдента, а 70 -из многомерного нормального распределения. При $r = 0$ - все наблюдения будут сгенерированы из многомерного нормального распределения, а при $r = 1$ все наблюдения будут сгенерированы из многомерного распределения Стьюдента.

В качестве генератора многомерного нормального распределения была использована функция `random.multivariate_normal()` библиотеки Numpy. Генератор многомерного распределения Стьюдента реализует описанные в равенстве 3 выражения с помощью функций `random.multivariate_normal()` [1] и `random.chisquarel()` [2] библиотеки Numpy.

[1] Документация: `https://numpy.org/doc/stable/reference/random/generated/numpy.random.multivariate_normal.html`

[2] Документация: `https://numpy.org/doc/stable/reference/random/generated/numpy.random.chisquare.html`

Рассмотрим более детально алгоритмы построения сетевых структур, а также специфику вычисления их статистической неопределённости.

Для построения максимального остовного дерева была использована функция `algorithms.tree.mst.maximum_spanning_tree()` библиотеки NetworkX [3]. Она реализует алгоритм Краскала[6] построение минимального остовного дерева, который итеративно строит остовное дерево, на каждом шаге присоединяя ребро наименьшего веса, добавление которого не вызовет появления цикла. Алгоритм может быть использован и для поиска максимального остовного дерева, выбирая на каждом шаге ребро наибольшего веса.

Остовное дерево, полученное как из эталонной, так и из выборочной сети, содержит N вершин и $M_1 = M_2 = N - 1$ рёбер. Так вершины соединяет только одно ребро, если существует ребёро (i, j), которое ошибочно включено в выборочную структуру($x_1^{ij} = 1$), то существует и ребро (k, s), котороя ошибочно не включено в структуру ($x_2^{ks} = 1$) и наоборот. Таким образом, ошибки I и II рода эквиваленты, а общая доля ошибок может быть найдена как

$$X = \frac{1}{2}\left(\frac{X_1}{M_1} + \frac{X_2}{M_2}\right) = \frac{X_1 + X_2}{2(N-1)} = \frac{X_1}{N-1} \tag{9}$$

Построение рыночного графа не потребовало использования никаких специальных алгоритмов. Рыночный граф легко получить, от-

[3] Документация: `https://networkx.github.io/documentation/stable/reference/algorithms/generated/networkx.algorithms.tree.mst.maximum_spanning_tree.html`

фильтровав список рёбер по весу. Для рыночного графа значения M_1 и M_2 по-прежнему константные, и равные $\binom{N}{2} - M$ и M соответственно, где M - количество рёбер в эталонном рыночном графе с некоторым выбранным значением порога θ, а $\binom{N}{2}$ - число всех возможных рёбер в графе с N вершинами. Общая доля ошибок X для рыночного графа равна:

$$X = \frac{1}{2}\left(\frac{X_1}{M_1} + \frac{X_2}{M_2}\right) = \frac{1}{2}\left(\frac{X_1}{\binom{N}{2} - M} + \frac{X_2}{M}\right) \tag{10}$$

Максимальная клика и максимальное независимое множество строятся на основе полученного рыночного графа. Так как таких структур в графе может быть несколько, для оценки неопределённости, мы будем искать максимальную клику с максимальным весом (maximum clique with maximal weight) и максимальное независимое множество с минимальным весом (maximum independent set with minimal weight).

Для поиска максимальной клики с максимальным весом в графе была использована функция `algorithms.clique.find_cliques()` библиотеки NetworkX[4], которая возвращает список всех клик графа, каждая из которых является кортежем из вершин графа, входящих в клику. Реализация поиска основана на алгоритме Брона — Кербоша[4] для поиска всех клик в неориентированном графе.Полученный список сортируется по размеру клик(числу элементов в кортеже), а

[4]Документация: `https://networkx.github.io/documentation/stable/reference/algorithms/generated/networkx.algorithms.clique.find_cliques.html`

потом из клик максимального размера выбирается клика с наибольшим весом. Вес клики определяется суммой весом рёбер рыночного графа, которые соединяют вершины, входящие в клику.

Так как максимальное независимое множество графа совпадает с максимальном кликой в обратном графе, поиск максимального независимого множество реализован через поиск максимальной клики в дополнении рыночного графа. Дополнением графа G является граф, в котором две вершины являются смежными, если они не смежны в исходном графе. Множество вершин у исходного графа и его дополнения совпадают, а объединённые множества рёбер обоих графов составляет множество рёбер полного графа на этом множестве вершин. Дополнение графа также называют обратным графом. Таким образом, чтобы найти максимальное независимое множество графа, можно применить описанный выше подход для поиска максимальной клики в дополнении рыночного графа, только вместо максимальной клики с наибольшим весом будем выбирать клику с наименьшим весом.

В отличие от максимального остовного дерева и рыночного графа значения M_1 уже не является константой, так как размер выборочной максимальной клики может изменяться. Теперь $M_1 = \binom{C_s}{2}$, а $M_2 = \binom{C_r}{2}$, где C_s - число веришн в выборочной максимальной клике, а C_r - число веришн в эталонной. Значение C_r - константа.

$$X = \frac{1}{2}\left(\frac{X_1}{M_1} + \frac{X_2}{M_2}\right) = \frac{1}{2}\left(\frac{X_1}{\binom{C_s}{2}} + \frac{X_2}{\frac{X_2}{\binom{C_r}{2}}}\right) \tag{11}$$

Алогоритм вычисления ошибок I и II рода (X_1 и X_2) одинаковый для каждой из структур. Он состоит в использовании функции `algorithms.operators.difference()` библиотеки NetworkX[5]. Эта функция принимает на вход два графа и возвращает граф с тем же множеством вершин и множеством рёбер, которые содержатся в первом графе, но не содержатся во втором. У полученного графа мы считаем количество рёбер. Таким образом, передавая в функцию выборочную структуру первым аргументов, а эталонную вторым, мы получаем ошибку I рода. Передавая структуры в обратном порядке, получаем ошибку II рода

4.1 Анализ эталонной сети

Данными для эскперимента являются 470 наблюдений за $N = 100$ акциями NASDAQ в период 2018-2019 годов. На основании этих данных, построим матрицу выборочной корреляции Пирсона $||\rho_{ij}||$ и матрицу вероятностей совпадения знаков, которые будут являться матрицами весов для эталонных сетей. К ним мы будем применять различные процедуры фильтрации и получать эталонные сетевые

[5]Документация: `https://networkx.github.io/documentation/stable/reference/algorithms/generated/networkx.algorithms.operators.binary.difference.html`

структуры, с помощью которых мы будем оценивать статистическую неопределённость.

Данные были загружены с помощью библиотеки Yahoo! Finance market data downloader[6]. Набор тикеров был получен путём веб-скрейпинга сайта `http://eoddata.com`.

Рассмотрим некоторые статистики эталонных сетевых структур с **корреляцией Пирсона** в качестве меры близости. В таблице 1 отображена зависимость числа рёбер от выбранного порога θ в эталонном рыночном графе.

θ	-1	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
|E|	4950	4584	3593	2282	1108	428	150	51	16

Таблица 1: Зависимость числа рёбер в рыночном графе от θ

В таблице 2 приведены размеры максимальной клики и максимального независимого множества в эталонном рыночном графе в зависимости от выбранного порога θ.

θ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
N clique	81	64	43	24	15	10	5	3
N mis	5	14	29	41	60	76	84	92

Таблица 2: Зависимость размеров структур от θ

[6] Документация: `https://pypi.org/project/yfinance/`

На рисунке 6 изображён эталонный рыночный граф с порогом $\theta = 0.5$, а также максимальная клики и максимальное независимое множество этого графа. Вершины, входящие в клику изображены синим цветом, входящие в независимое множество - красным.

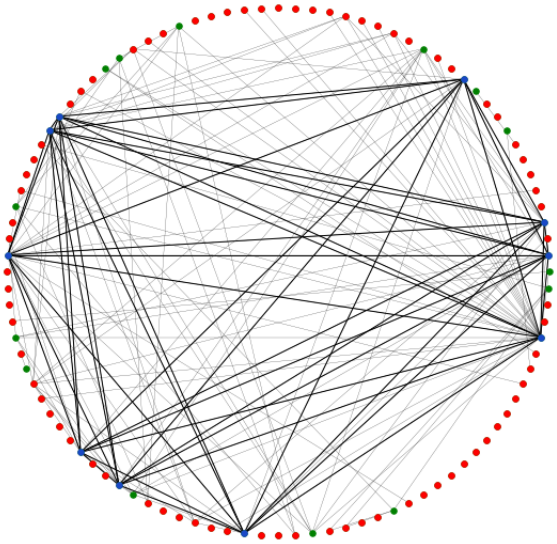


Рис. 6: Эталонный рыночный граф с порогом $\theta = 0.5$

Теперь рассмотрим те же статистики эталонных сетевых структур, но с **вероятностями совпадения знаков** в качестве меры близости. Порог θ преобразован в θ_γ согласно (6).

В таблицах 3 и 4 отображены зависимость числа рёбер от выбранного порога и размеры максимальной клики и максимального независимого множества.

θ	-1	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
θ_γ	0	0.5	0.53	0.56	0.6	0.63	0.67	0.7	0.75	0.8	0.86
|E|	4950	3906	3299	2341	1299	554	226	83	26	10	2

Таблица 3: Зависимость числа рёбер в рыночном графе от θ и θ_γ

θ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
θ_γ	0.5	0.53	0.56	0.6	0.63	0.67	0.7	0.75	0.8	0.86
N clique	73	55	37	25	16	10	5	4	3	2
N mis	14	18	29	38	52	68	80	88	95	98

Таблица 4: Зависимость размеров структур от θ и θ_γ

4.2 Измерение неопределённости сетевых структур

Для каждой из структур будем применять две меры неопределённости: корреляция Пирсона и вероятность совпадения знаков.

Для корреляции Пирсона основным экспериментом будет измерение общей доли ошибки в зависимости от числа наблюдений n для различых параметром r. Таким образом, мы сможем изучить, как влияет на статистическую неопределённость отклонения от нормального распределения. Мы увидим, что чем ближе значения параметра r к единице, то есть чем больше отклонение, тем больше общая доля ошибок для каждой из структур.

Для вероятности совпадения знаков мы рассмотрим зависимость общей доли ошибки от параметра r для различных значений параметра r для различных значений числа наблюдений n. Так мы изучим влияние отклонения от нормального распределения на статистическую неопределённость структуры, построенной на вероятности совпадения знаков. Мы увидим, что подобного отклонения практически нет для любой из выбранных структур. Кроме того, мы сможем исследовать величину общей доли ошибки в зависимости от выбранной меры близости для выбранного значений n.

Доли ошибок для корреляции знаков мы сравним с ошибками для корреляции Пирсона при $r = 1$ и $r = 0$ как для случаев с наибольшей и наименьшей ошибками соответственно. Мы будем сравнивать с единственным значением ошибки для знаков, так как при изменении r значение общей доли ошиюки будет оставаться практически неизменной. Для максимальной клики и максимального независимого множества помимо общей доли ошибки, будут также представлены доли ошибок I и II рода.

4.2.1 Максимальное остовное дерево

На рисунке 7 представлен график зависимости общей доли ошибки от числа наблюдений для различных r. При $r = 0$, то есть при генерации только из нормального распределения, порог $\mathcal{E}_0 = 0.1$ достигается примерно при $n = 4000$ наблюдений, далее ошибка про-

должает медленно убывать. В работе [7] порог достигался при более чем 10000 набледедий.

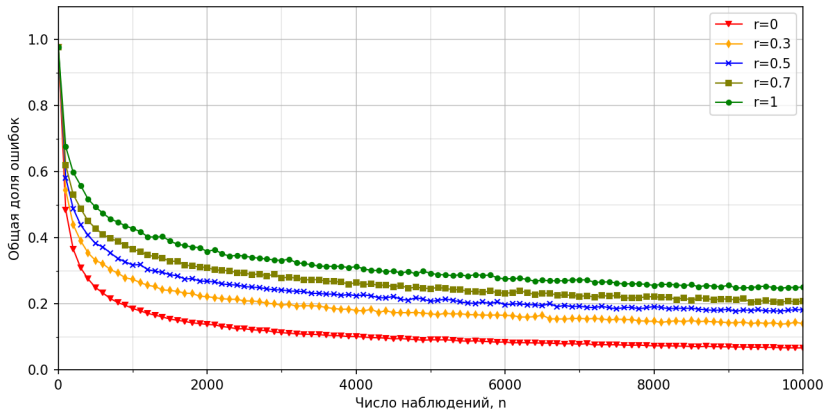


Рис. 7: Зависимость общей доли ошибок от n, MST

На рисунке 8 представлен график зависимости общей доли ошибки от параметра r при различных n. Уже при $r > 0.1$ числа наблюдений $n = 10000$ не хватает, чтобы достичь порога $\mathcal{E}_0 = 0.1$.

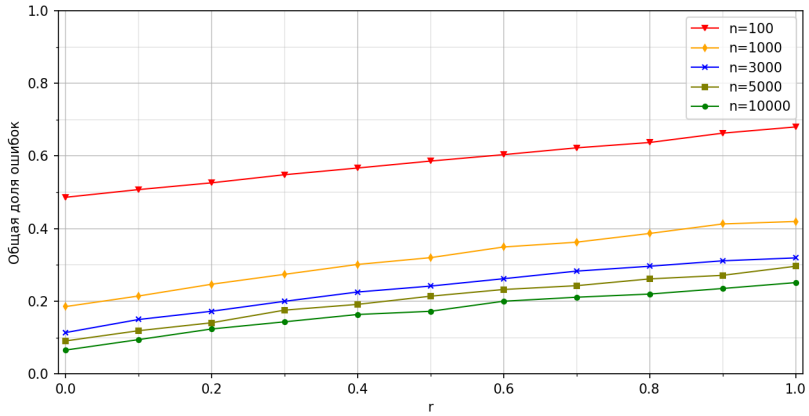


Рис. 8: Зависимость общей доли ошибок от r, MST

Теперь рассмотрим вероятность совпадения знаков в качестве меры близости. На рисунке 9 представлен график зависимости общей доли ошибки от параметра r при различных n. Видно, что изменения в распределении практически никак не влияют на значение ошибки. Однако достичь порога $\mathcal{E}_0 = 0.1$ не получается даже при $n = 10000$. Судя по графику, значение общей доли ошибок стремится к 0.5 при увеличении числа наблюдений n. Подобная зависимость общей доли ошибки от параметра r будет наблюдаться и для остальных структур, из чего мы сможем сделать вывод, что вероятность совпадения знаков устойчива к отклонениям наблюдений от нормального распределения подобного рода.

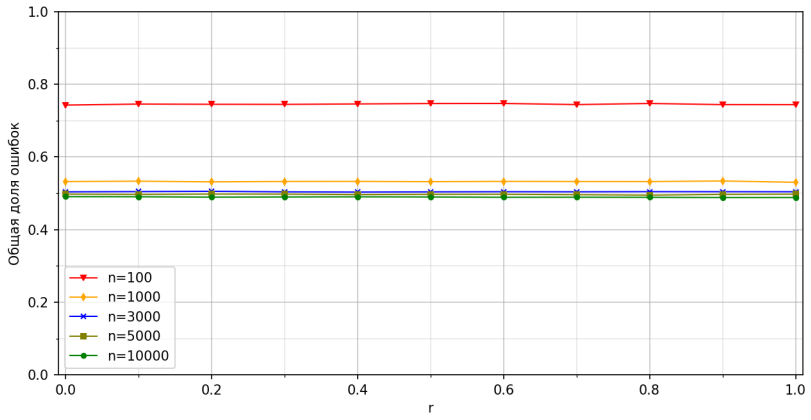


Рис. 9: Зависимость общей доли ошибок от r для знаков, MST

На слудеющем рисунке 10 представлены графики сравнения неопределённости при использовании коррелиции Пирсона и вероятности выпадения знаков. Слева изображён график зависимости общей до-

ли ошибки от параметра r при различных n для корреляции Пирсона и корреляции вероятности совпадения знаков. Справа - график, показывающий значения ошибки для корреляции Пирсона при $r = 0$ и $r = 1$ и для вероятности совпадения знаков, общая доля ошибок для которой сохраняется при изменении r. Видно, что даже при $r = 1$, структура на основе корреляции Пирсона более стабильна.

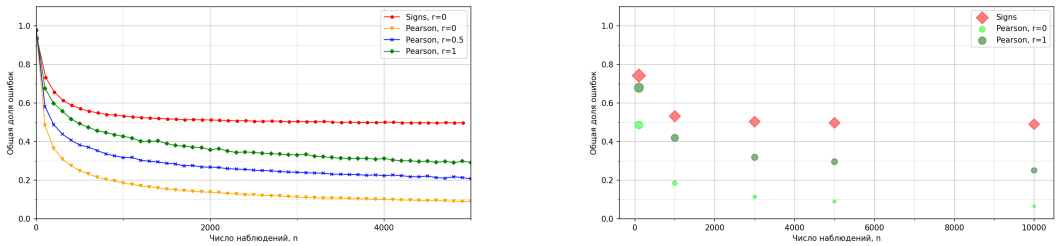


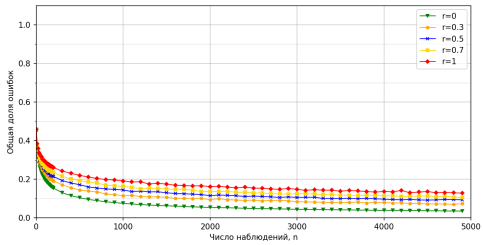
Рис. 10: Зависимость общей доли ошибок от n, MST

4.2.2 Рыночный граф

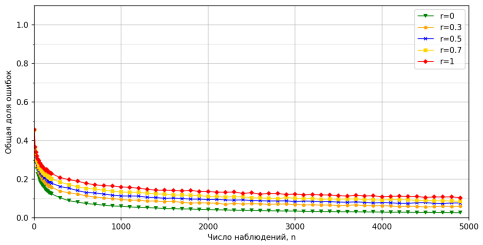
В отличие от максимального остовного дерева рыночный граф имеет дополнительный параметр: порог θ. Мы рассмотрим несколько значений порга θ: $\theta \in [0.1, 0.3, 0.5, 0.7]$. Каждый график будет соответствовать своему значению θ.

На рисунке 11 представлен график зависимости общей доли ошибки от числа наблюдений для различных r при различных значениях порога θ. В отличие от максимального остовного дерева, уровень $\mathcal{E}_0 = 0.1$ достигается при гораздо меньшем числе наблюдений. При пороге выше $\theta = 0.5$ уровень $\mathcal{E}_0 = 0.1$ достигается для любого зна-

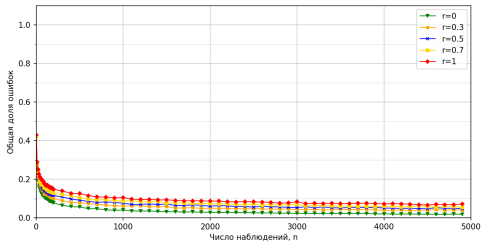
чения r.



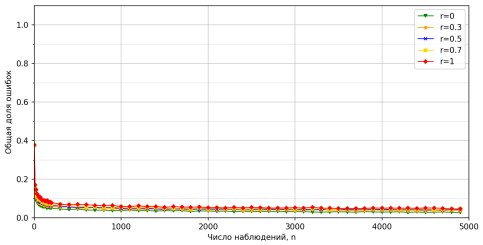
(a) $\theta = 0.1$



(b) $\theta = 0.3$



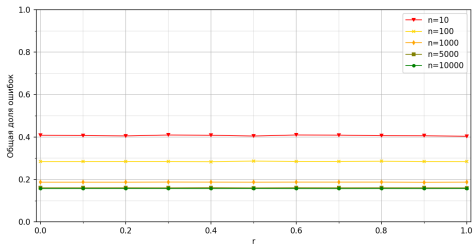
(c) $\theta = 0.5$



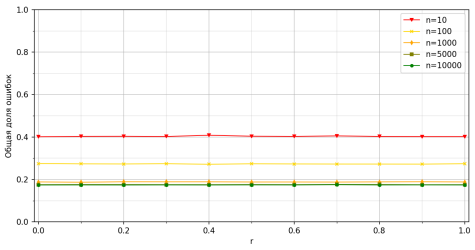
(d) $\theta = 0.7$

Рис. 11: Зависимость общей доли ошибок от n, MG

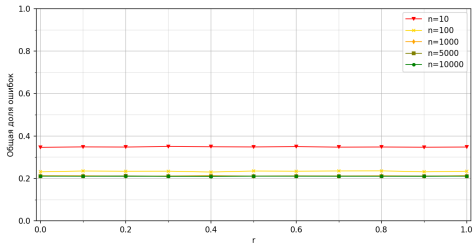
Для меры близости на основе вероятности совпадения знаков на рисунке 12 представлен график зависимости общей доли ошибки от параметра r при различных n. Как и для максимального остовного дерева, изменения в распределении также практически никак не влияют на значение ошибок. Достичь порога $\mathcal{E}_0 = 0.1$ всё также не удаётся, хотя общая доля ошибка меньше: около $\mathcal{E} = 0.2$ для всех рассмотренных значений порога θ. Согласно условию (6) при работе со знаками, мы преобразуем порог θ в θ_γ. Значения θ_γ также представлено.



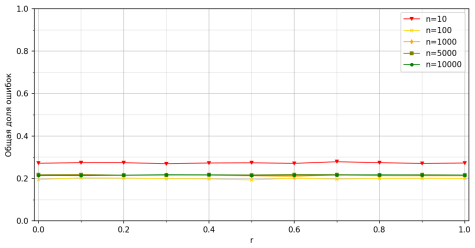
(a) $\theta = 0.1, \theta_\gamma = 0.53$



(b) $\theta = 0.3, \theta_\gamma = 0.6$



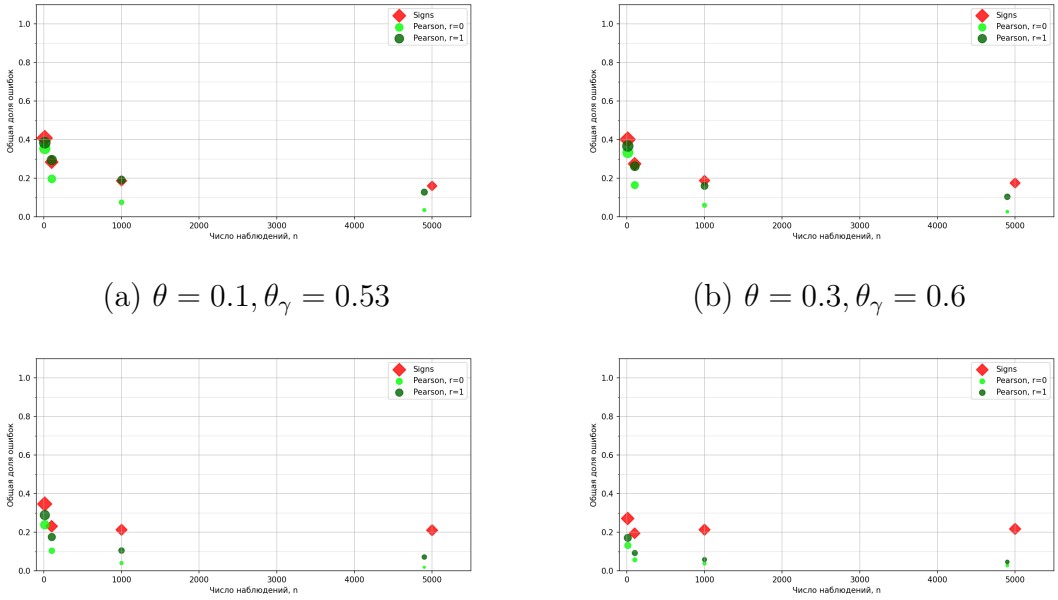
(c) $\theta = 0.5, \theta_\gamma = 0.67$



(d) $\theta = 0.7, \theta_\gamma = 0.75s$

Рис. 12: Зависимость общей доли ошибок от r для знаков, MG

На слудеющем рисунке 13 представлены график сравнения ошибок при использовании корреляции Пирсона и корреляции вероятности совпадения знаков. Заметим, что при маленьких значениях θ ошибки для корреляции Пирсона при $r = 1$ и корреляции знаков практически идентичны при разных n, однако при увеличии порого θ значения ошибки для корреляции Пирсона становится меньше. Это может быть вызвано преобразованием порога, либо тем, что знаки показывают себя лучше при использовании с маленьким θ, то есть при большом количстве рёбер в сети.



(a) $\theta = 0.1, \theta_\gamma = 0.53$ (b) $\theta = 0.3, \theta_\gamma = 0.6$

(c) $\theta = 0.5, \theta_\gamma = 0.67$ (d) $\theta = 0.7, \theta_\gamma = 0.75$

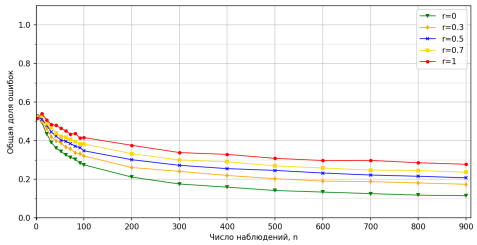
Рис. 13: Сравнение ошибки для обеих мер, MG

4.2.3 Максимальная клика

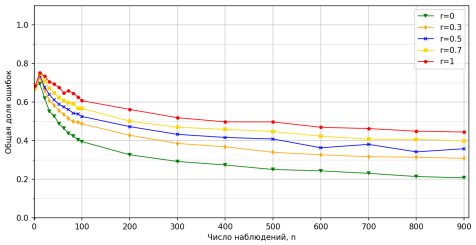
Максимальная клика и максимальное назависимое множество строятся на основе рыночного графа. Порог θ для максимальной клики будем брать из множества $\theta \in [0.1, 0.3, 0.5, 0.7]$.

На рисунке 14 изображены графики зависимости общей доли ошибок от количества наблюдений n для различных значений порога и, соответственно, для различных размеров клики. Значение общей доли ошибок остаётся достаточно большим, в сравнении со значениями ошибки в рыночном графе. Оно варьируется от $\mathcal{E} = 0.2$ до $\mathcal{E} = 0.4$ в зависимости от значения r. При большом и маленьком

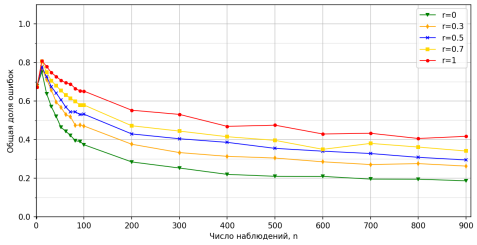
порогах $\theta = 0.7$ и $\theta = 0.1$ значения ошибок меньше, чем при средних $\theta = 0.3$ и $\theta = 0.5$. При $\theta = 0.7$ и $\theta = 0.1$ достигается значение порога $\mathcal{E}_0 = 0.1$



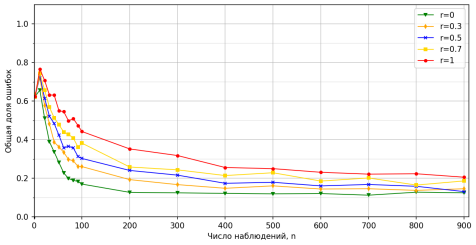
(a) $\theta = 0.1, N = 64$



(b) $\theta = 0.3, N = 24$



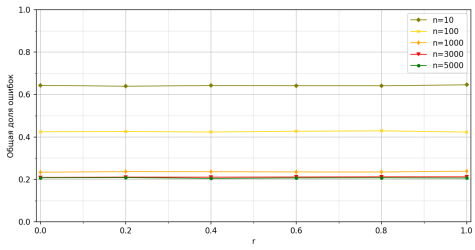
(c) $\theta = 0.5, N = 10$



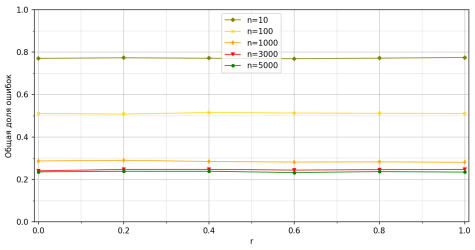
(d) $\theta = 0.7, N = 3$

Рис. 14: Зависимость общей доли ошибок от n, MC

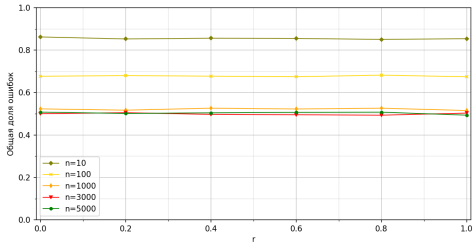
Для меры близости на основе вероятности совпадения знаков на рисунке на рисунке 15 представлен график зависимости общей доли ошибки от параметра r при различных n. Также как и для корреляции Пирсона, ошибка больше при средних значениях порога θ.



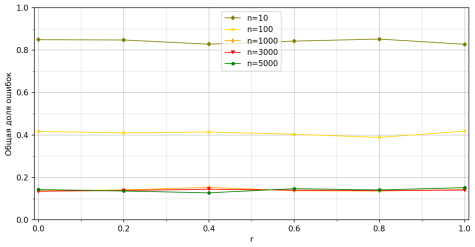
(a) $\theta = 0.1, \theta_\gamma = 0.53, N = 55$



(b) $\theta = 0.3, \theta_\gamma = 0.6, N = 25$



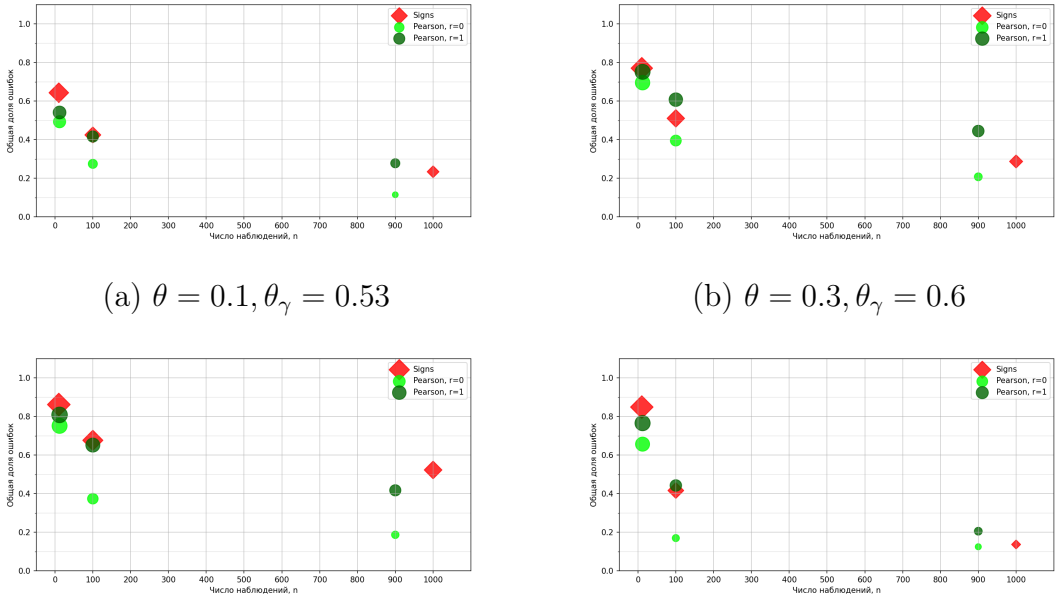
(c) $\theta = 0.5, \theta_\gamma = 0.67, N = 10$



(d) $\theta = 0.7, \theta_\gamma = 0.75, N = 4$

Рис. 15: Зависимость общей доли ошибок от r для знаков, MC

На графике 16 сравниваются значения ошибок при использовании корреляции Пирсона и корреляции вероятности совпадения знаков. Заметим, что для значений $100 < n < 1000$ ошибка для корреляции Пирсона и $r = 1$ равна примерно равна или даже превосходит ошибку при использовании вероятностей совпадения знаков.



(a) $\theta = 0.1, \theta_\gamma = 0.53$

(b) $\theta = 0.3, \theta_\gamma = 0.6$

(c) $\theta = 0.5, \theta_\gamma = 0.67$

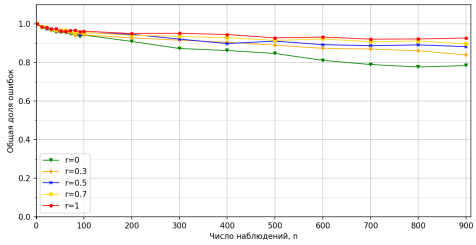
(d) $\theta = 0.7, \theta_\gamma = 0.75$

Рис. 16: Сравнение ошибки для обеих мер, MC

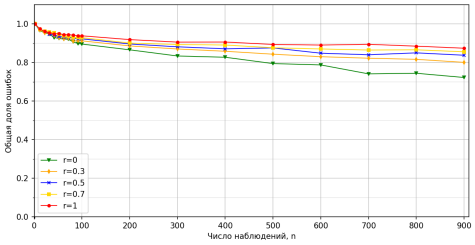
4.2.4 Максимальное независимое множество

По аналогии с максимальной кликой, мы будем рассматривать значения общей доли ошибок для разных порогов. Пороги для максимальное независимое множества будут иными: $\theta \in [0, 0.05, 0.1, 0.15]$.

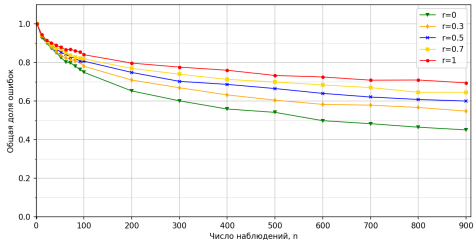
На рисунке 17 изображены графики зависимость общей доли ошибки от количества наблюдений n для различных значений порога. Небольшое увеличение порога приводит к существенному уменьшению ошибки. С ростом числа наблюдений ошибка уменьшается существенно медленне, чем у рассмотренной выше максимальной клики, и не достагает порога $\mathcal{E}_0 = 0.1$



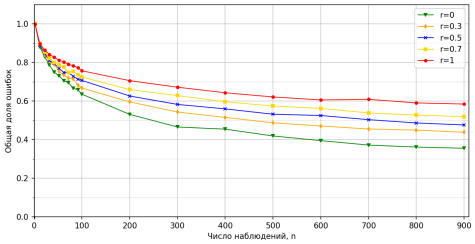
(a) $\theta = 0, N = 5$



(b) $\theta = 0.05, N = 8$



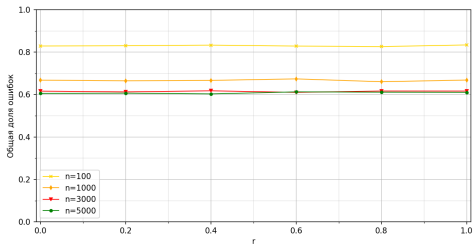
(c) $\theta = 0.1, N = 14$



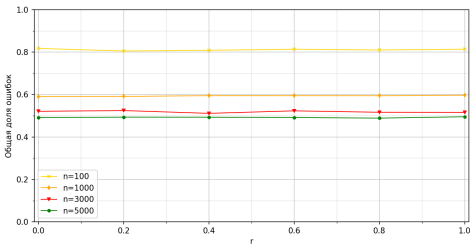
(d) $\theta = 0.15, N = 21$

Рис. 17: Зависимость общей доли ошибок от n, MIS

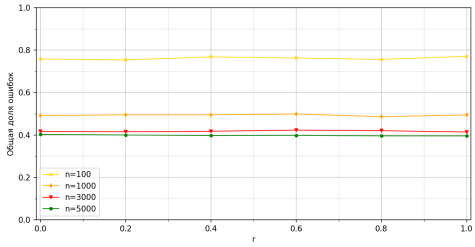
На следующем графике 18 представлена зависимость общей доли ошибок от параметра r с вероятностью совпадения знаков в качестве меры близости. Наблюдается схожая с корреляцией Пирсона тенденция с существенном уменьшение ошибки при небольшом увеличении θ.



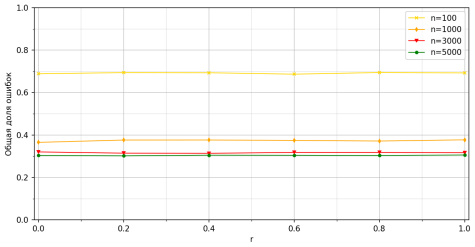
(a) $\theta = 0.05, \theta_\gamma = 0.51, N = 17$



(b) $\theta = 0.1, \theta_\gamma = 0.53, N = 18$



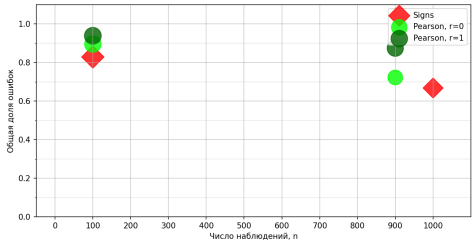
(c) $\theta = 0.15, \theta_\gamma = 0.55, N = 23$



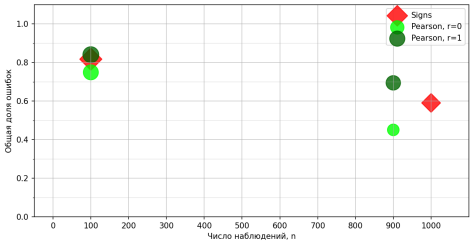
(d) $\theta = 0.2, \theta_\gamma = 0.56, N = 29$

Рис. 18: Зависимость общей доли ошибок от n, MIS

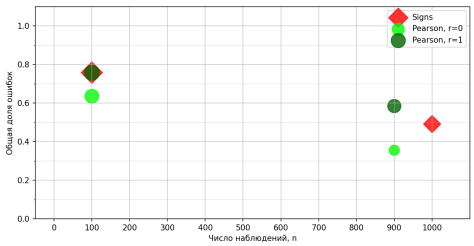
Сравнение ошибок для обеих мер представлено на график 19. Видно, мера вероятности совпадения знаков проявляет себя немного лучше, чем мера корреляции Пирсона.



(a) $\theta = 0.05, \theta_\gamma = 0.51$



(b) $\theta = 0.1, \theta_\gamma = 0.53$



(c) $\theta = 0.15, \theta_\gamma = 0.55$

Рис. 19: Сравнение ошибки для обеих мер, MIS

5 Заключение

В данной работе была изучена статистическая неопределённость различных сетевых структур на основе $\mathcal{E}$-меры неопределённости, предложенной в работе[7]. Предполагая, что некоторая часть наблюдений за доходностями акций распределено не нормально, а имеет $t-$распределение, или распределение Стьюдента, мы изучили поведение неопределённости структур, для который корреляция Пирсона использована в качестве меры близости. В результате серии экспериментов оказалось, что чем большая доля наблюдений имеет рас-

пределение Стьюдента, тем больше неопределённость структуры, то есть тем хуже корреляция Пирсона отражает взаимосвязь между слуайными величинами. Для тех же структур, чья мера близости основывалась на вероятности совпадения знаков, было показано экспериментально, что доля наблюдений, имеющая распределение Стьюдента, не изменяет значение статистической неопределённости. Другими словами, мера близости, основанная на вероятности совпадения знаков гораздо менее чувствительна к распределению случайных величин.

Кроме того, было проведено сравнение значений мер статистической неопределённости структур, основанных на корреляции Пирсона и на вероятности совпадения знаков. Предполагалось, что использование корреляции вероятности совпадения знаков в качестве меры близости позволит уменьшить показатели неопределённости структур, особенно при высокой доли наблюдений, имеющих распределение Стьюдента. Однако, эскперименты показали, что меры неопределённости, построенные на корреляции Пирсона практически всегда имеют меньшую неопределённость, даже если все наблюдения имеют распределение Стьюдента. В этом случае неопределённость структур, построенных на вероятности совпадения знаков, может быть немного меньше или равна.

При сравнении результатов, полученных в этой работе и результатов представленных в работе[7], возник вопрос о влиянии началь-

ных данных на оценку статистической неопределённости структур. Так, в работе[7] мера неопределённости для максимального остовного дерева не достигала порога $\mathcal{E}_0 = 0.1$ даже при 10000 наблюдениях, тогда как в ходе экспериментов в рамках данной работы, удалось достигнуть порога уже при 4000 наблюдениях. С другой стороны, непределённость таких сетевых структур как максимальная клика и максимальное независимое множество, оказалась гораздо выше в данной работе, чем в работе [7] для аналогичных значений n, хотя значения для рыночного графа сопоставимые. Тренд особой зависимости непределённости рыночного графа от его порога θ, который предполагает наибольшее значение неопределённости при средних значениях θ и меньшее при больших $\theta > 0.7$ или малых $\theta < 0.2$ сохраняется и в экспериментах в данной работе. При этом его можно наблюдать как при использовании корреляции Пирсона, так и при использовании вероятности совпадения знаков в качестве меры близости.

Весь исходный код для построения сетевых структур, а также для генерации наблюдений и измерения меры неопределённости доступен в репозитории на github: `https://github.com/ArtamonovDen/measures-of-uncertainty`.

Список литературы

[1] Grigory Bautin и др. “Simple measure of similarity for the market graph construction”. в: *Computational Management Science* 10 (июнь 2013). DOI: 10.1007/s10287-013-0169-3.

[2] Vladimir Boginski, Sergiy Butenko и Panos M. Pardalos. *On Structural Properties of the Market Graph*. 2003.

[3] Z. I. Botev и P. L'Ecuyer. “Efficient probability estimation and simulation of the truncated multivariate student-t distribution”. в: *2015 Winter Simulation Conference (WSC)*. 2015, с. 380—391.

[4] Coen Bron и Joep Kerbosch. “Algorithm 457: Finding All Cliques of an Undirected Graph”. в: *Commun. ACM* 16.9 (сент. 1973), с. 575—577. ISSN: 0001-0782. DOI: 10.1145/362342.362367. URL: https://doi.org/10.1145/362342.362367.

[5] Wei-Qiang Huang, Xin-Tian Zhuang и Shuang Yao. “A network analysis of the Chinese stock market”. в: *Physica A-statistical Mechanics and Its Applications - PHYSICA A* 388 (июль 2009), с. 2956—2964. DOI: 10.1016/j.physa.2009.03.028.

[6] Kruskal J.B. “On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem”. в: *Proceedings of the American Mathematical Society* 7 (1956), с. 48—50.

[7] V. A. Kalyagin и др. “Measures of uncertainty in market network analysis”. в: (2013). DOI: 10.1016/j.physa.2014.06.054. eprint: arXiv:1311.2273.

[8] Rosario N. Mantegna. “Hierarchical Structure in Financial Markets”. в: (1998). DOI: 10.1007/s100510050929. eprint: arXiv:cond-mat/9802256.

[9] Rosario Mantegna и H. Stanley. *An Introduction to Econophysics: Correlations and Complexity in Finance*. т. 53. дек. 2000. DOI: 10.1063/1.1341926.

[10] Bilal Memon, Hong Yao и Rabia Tahir. “General election effect on the network topology of Pakistan’s stock market: network-based study of a political event”. в: *Financial Innovation* 6 (янв. 2020). DOI: 10.1186/s40854-019-0165-x.

[11] Eder Pereira и др. “Multiscale network for 20 stock markets using DCCA”. в: *Physica A: Statistical Mechanics and its Applications* (май 2019), с. 121542. DOI: 10.1016/j.physa.2019.121542.

[12] Anderson TW. *An introduction to multivariate statistical analysis*. 1957.

[13] Arseniy Vizgunov и др. “Network approach for the Russian stock market”. в: *Computational Management Science* 11 (апр. 2014). DOI: 10.1007/s10287-013-0165-7.

[14] Gang-Jin Wang, Chi Xie и Shou Chen. “Multiscale correlation networks analysis of the US stock market: A wavelet analysis”. в: *Journal of Economic Interaction and Coordination* 12 (окт. 2017), с. 561—594. DOI: 10.1007/s11403-016-0176-x.