Конспект лекции 7

Юдин Е.Б., Юдина М.Н.

Сетевые мотивы

# Сетевые мотивы

Здравствуйте, я представляю лекцию 1 “Сетевые мотивы” по дисциплине “Модели больших сетевых структур и сетевые процессы”.

Разработчиками курса являются сотрудники Омского государственного технического университета Юдин Е.Б., Юдина М.Н., Бадрызлов В.А., а также сотрудник института математики им. С.Л. Соболева Логинов К.К.

Данную лекцию разработала Юдина М.Н. в соавторстве с Юдиным Е.Б.



# Исследование сетей

Подход, заключающийся в выявлении в графах сетей так называемых сетевых мотивов, востребован при исследовании действительно больших сетей, структура которых кажется запутанной и сложной: Интернет, социальные сети, сети белковых взаимодействий и т.д.



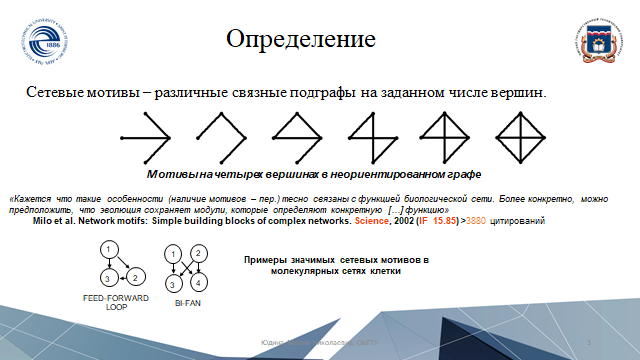
# Определение

Этот подход заключается в поиске подграфов небольшого размера, которые могут встречаться в исследуемой сети чаще, чем этого можно было бы ожидать.

Понятие сетевых мотивов впервые было использовано для анализа молекулярных сетей клетки. Майло объясняет наличие сетевых мотивов тем, что «эволюция сохраняет модули, которые выполняют конкретную … функцию». «… evolution preserves modules that define specific … functions»

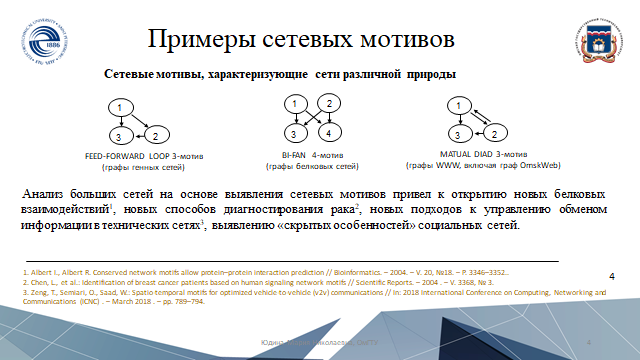
Сетевые мотивы на k-вершинах выявляются путем сравнения частот встречаемости связных подграфов в исследуемом графе и в его рандомизированных версиях.

Те подграфы, которые встречаются в исследуемом графе чаще, чем в рандомизированных версиях, называются сетевыми мотивами.



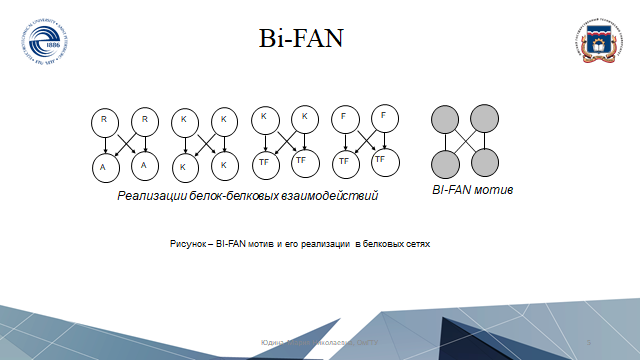
# Примеры сетевых мотивов

Анализ сетевых мотивов в больших сетях привел к открытию новых белковых взаимодействий, новых способов диагностирования рака, новых подходов к оценке надежности сетей, определению «скрытых» особенностей сетей.



# Bi-FAN

Почему Bi-FAN характерен для белок-белковых взаимодействий. Дело в том, что в сети белковых взаимодействий часто встречаются рецепторы (R), взаимодействующие с субстратом (A), такие как G-белковые рецепторы, и рецепторы, взаимодействующие с G-белком (GPCR-рецепторы), в белковых сетях также часто встречаются каскады с участием киназ (K), фермента фосфатазы (F) и различных факторов транскрипции (TF).



# FEED-FORWARD LOOP

Что касается Сетевого мотива FEED-FORWARD LOOP, то этот мотив характерен для генных сетей. Ведь этот сетевой мотив определяет взаимодействие, при котором целевой ген В регулируется генами A и Б, при этом ген Б также регулируется геном A. Наличие значимого сетевого мотива FEED-FORWARD LOOP объясняется особенностями использования факторов транскрипции. Процесс синтеза белка исследуемого гена зависит от присутствия белков (факторов транскрипции), синтезируемых на основе других генов. Обычно необходимо присутствие целого ряда различных факторов транскрипции.



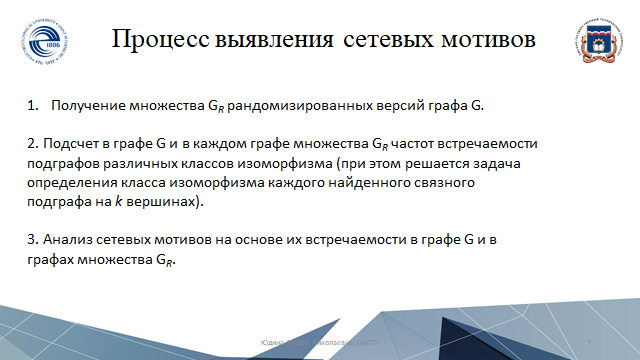
# Этапы выявления сетевых мотивов

Задачу нахождения сетевых мотивов на *k* вершинах (*k*-мотивов) в графе G можно разбить на три подзадачи.

1. Получение множества G*R* рандомизированных версий графа G*.*

2. Подсчет в графе G и в каждом графе множества G*R* частот встречаемости подграфов различных классов изоморфизма (при этом решается задача определения класса изоморфизма каждого найденного связного подграфа на *k* вершинах).

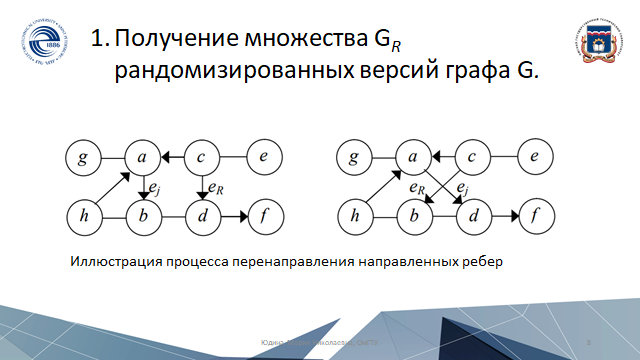
3. Анализ сетевых мотивов на основе их встречаемости в графе G и в графах множества G*R*.



# Получение множества GR рандомизированных версий графа G.

Для получения множества рандомизированных графов, как правило, используются специальные процедуры рандомизации. Рандомизация графа начинается с создания графа-копии исследуемого графа G, затем выполняется перенаправление ребер/дуг этого графа. Для каждого ребра/дуги делаются попытки перенаправления концаребра/дуги, инцидентного вершине *u*,на другую вершину. По некоторым причинам перенаправление ребер/дуг может не выполниться (например, если оба ребра начинаются с одной и той же вершины). Тогда случайный выбор повторяется до тех пор, пока обмен не осуществится или количество попыток обмена не превысит заданное пользователем значение.

На рисунке, два направленных ребра e\_j ={a, b} и e\_R={c, d} графа выбираются случайно. Сначала ребра e\_j ={a, b} и e\_R={c, d} удаляются из графа, а потом формируются новые ребра с теми же именами. Это происходит следующим образом. Вершина a связывается с вершиной d (формируя e\_j), а затем вершина d связывается с вершиной c (формируя e\_R). Перед осуществлением такого перенаправления необходимо убедиться, что между вершинами a и d, а также между вершинами c и b, до перенаправления ребер не существовало. При этом у каждой вершины число входящих или исходящих дуг не меняется.

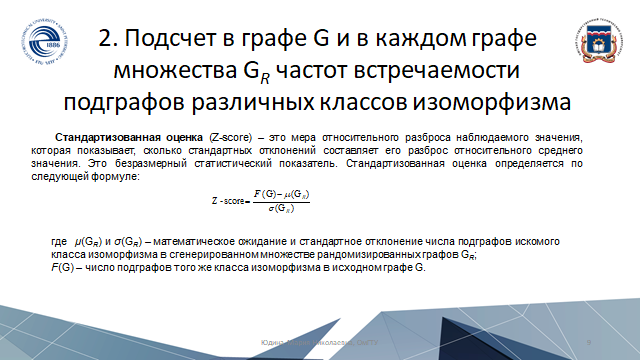


# Подсчет в графе G и в каждом графе множества GR частот встречаемости подграфов различных классов изоморфизма

**Стандартизованная оценка** (Z-score) – это мера относительного разброса наблюдаемого значения, которая показывает, сколько стандартных отклонений составляет его разброс относительного среднего значения. Это безразмерный статистический показатель. Стандартизованная оценка определяется по формуле, представленной на слайде

А именно , берется разность *F*(G) – числа подграфов того же класса изоморфизма в исходном графе G и математическое ожидание числа подграфов искомого класса изоморфизма в сгенерированном множестве рандомизированных графовG*R* . И эта разнасть нормируется  на *σ*(G*R*) –стандартное отклонение числа подграфов искомого класса изоморфизма в сгенерированном множестве рандомизированных графовG*R*;

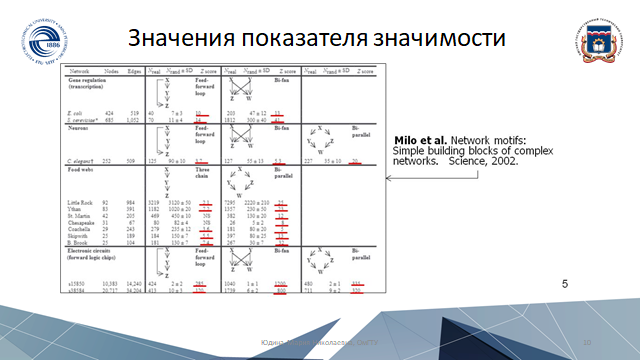
Сетевым мотивами считаются такие подграфы, которые относятся к классам изоморфизма с показателем Z-score превышающее некоторое пороговое значение («по умолчанию» в программе FANMOD используется пороговое Z-score = 2).



# Значения показателя значимости

На рисунке подчеркнуты те значения Z-Score, которые больше двух для соответсвующего подграфа.

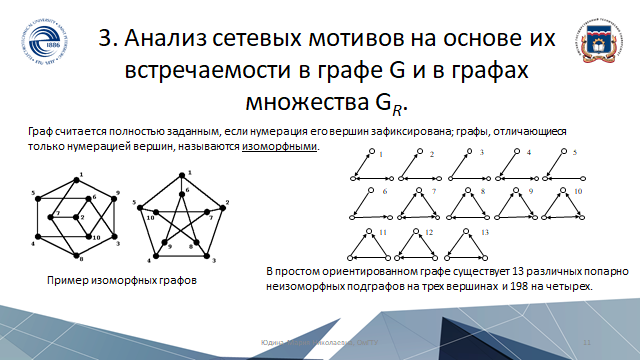
Для оценки значимости сетевых мотивов известны и другие статистические показатели, например P-value и показатель Пикарда



# Анализ сетевых мотивов на основе их встречаемости в графе G и в графах множества GR.

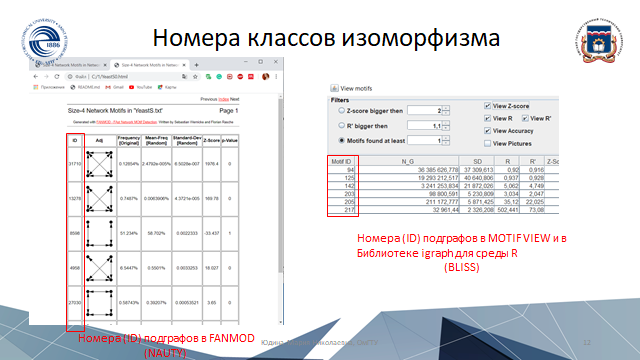
Теперь что касается третьей подзадачи при выявлении сетевых мотивов.

Конечно, подсчёт частот встречаемости подграфов задача сложная, но мы о ней поговорим чуть позже. Пока же заметим, что каждый раз при обнаружении очередного подграфа нужно определить к какому классу изоморфизма он относится. С этой целью в задачах выявления сетевых мотивов используются алгоритмы, основанные на применении канонической нумерации графов. Каноническая нумерация графов позволяет пронумеровать все возможные графы так, что одно и то же целое число будет соответствовать графам тогда и только тогда, когда они изоморфны. Таким образом, каждый из подграфов на заданном числе вершин будет соответствовать номеру одного из сетевых мотивов на том же числе вершин. При этом существуют различные системы нумерации и алгоритмы получения номеров. В частности, в самой популярной программе для расчёта мотивов FANMOD используется алгоритм Nauty и одноименная нумерация. В пакете igraph используется каноническая нумерация BLISS. В WolframMathematica используется нумерация Пиперно. Ряд специализированных программ использует нецифровую систему нумерации для решения задачи распознавания химических соединений и обеспечения их ускоренного поиска в специальных базах данных.



# Номера классов изоморфизма

В частности, в самой популярной программе для расчёта мотивов FANMOD используется алгоритм Nauty и одноименная нумерация. В пакете igraph используется каноническая нумерация BLISS. В WolframMathematica используется нумерация Пиперно. Ряд специализированных программ использует нецифровую систему нумерации для решения задачи распознавания химических соединений и обеспечения их ускоренного поиска в специальных базах данных



# Частоты встречаемости подграфов, сама по себе, важная структурная характеристика

Следует также заметить, что частоты встречаемости сетевых мотивов даже без анализа их значимости, т.е. сравнение с рандомизированными версиями сетей, являются важной структурной **характеристикой**, некоторым специфическим рисунком сети, отражающем микровзаимодействие её элементов.

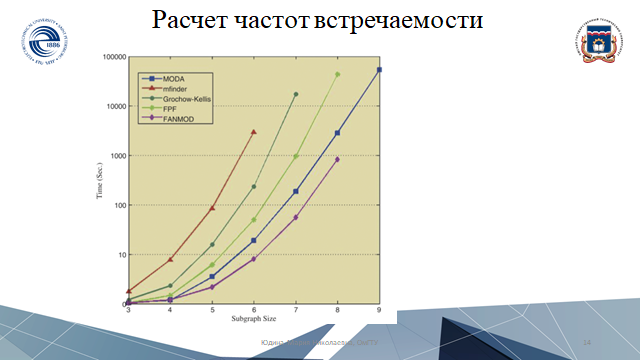
Рисунке изображены два графа эго-сетей социальной сети «Вконтакте». Пунктирными линиями обозначены ребра, связывающие вершины, которым соответствуют исследуемые узлы, с ближайшим окружением этих вершин, а сплошными – ребра, связывающие вершины из ближайшего окружения между собой. Один из графов эго-сетей (слева) описывает ближайшее окружение вершины, которая соответствует студенту ОмГТУ (назовем это узел «Студент»), а другая – преподавателю ОмГТУ (назовем этот узел «Преподаватель»).

Из Таблицы справа, в свою очередь, даже несмотря на рисунок, можно также видеть, что в окрестности вершины, соответствующей узлу «Преподаватель», доля подгафов на четырех вершинах с большим числом ребер значительно больше, чем в окрестности вершины, соответствующей узлу «Студент»,  при этом из анализа частот подграфов на трех вершинах видно, что сеть Студента значительно плотнее (83% треугольников). Такой результат является следствием того, что узлы разных сообществ эго-сети узла «Студент» не взаимодействуют между собой.



# Время расчета частот встречаемости

Выявление сетевых мотивов – это достаточно трудоемкий процесс, а самой трудоемкой его частью является расчет частот встречаемости подграфов.



# Классификация методов расчета подграфов: полный перебор

Метод Верника – это метод посл. перебора всех подграфов. Подграфы перебираются путем построения дерева подграфов, на заданном ярусе которого находятся все подграфы, имеющие соответствующий размер: вершины, ребра, 3-мотивы, можно расширить до 4-мотивов и т.д.

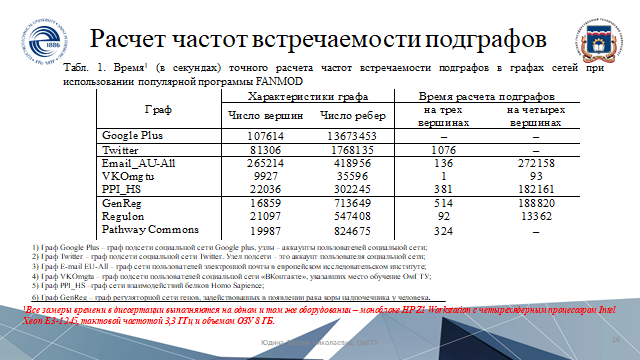


# Расчет частот встречаемости подграфов

Выявление сетевых мотивов – это достаточно трудоемкий процесс, и самой трудоемкой его частью является расчет частот встречаемости подграфов. Точный расчет частот встречаемости подграфов был приемлем в 2002 году, когда была предложена концепция использования сетевых мотивов. Тогда изучались относительно небольшие графы, но сейчас, например, сети известных белковых взаимодействий содержат уже не тысячи вершин, а сотни тысяч вершин и ребер. А для исследования социальных сетей точный расчет вообще становится практически невозможен.

На слайде также представлены данные о времени, которое требуется для точного расчета частот встречаемости подграфов на трех и четырех вершинах. Иногда это несколько суток работы. Прочерки означают, что за 5 дней работы не удалось осуществить расчет. Использовалась самая известная программа, используемая для расчетов больших сетей – программа Fanmod.

При этом использовался алгоритм Верника, введенный на предыдущем слайде



# Классификация методов расчета подграфов: статистические методы

Поэтому стали разрабатываться методы, позволяющие получать статистические оценки для частот встречаемости подграфов. Причем ряд методов позволяет работать только с неориентированными графами. Часть – получать только относительные частоты встречаемость подграфов

Например, методы случайного  блуждания,  выделены желтым, такие как ESA – метод случайного выбора ребра и расширения его до подграфа необходимого размера, и другие, остованные на построении цепи Маркова между подграфами, эти методы позволяют рассчитать только относительные частоты встречаемости, концентрацию подграфов.

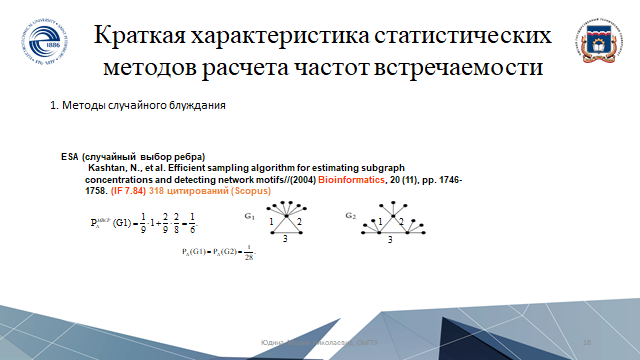
Методы перечисления подграфов со случайным отсечением выборки, обозначены зеленым фактически реализуют перебор подграфов, просто не все подграфы исследуются.

Идея методов обобщений графов, обозначенных голубым цветом, заключается в точном расчете частот встречаемости подграфов в небольшой части графа или в редуцированном графе, полученном из исходного графа. В них слишком много эвристики.



# Краткая характеристика статистических методов расчета частот встречаемости

Среди известных стат. Методов случайного блуждания, можно выделить метод случайного выбора ребра. Исторически это был первый статистический метод для расчёта встречаемости подграфов . Метод случайного выбора ребра основан на реализации случайного процесса, который начинается с равновероятного выбора ребра. Далее из ближайшего окружения вершин, инцидентных выбранному ребру, выбирается третья вершина. Потом из ближайшего окружения этих трех вершин – четвертая и т.д. до получения реализации подграфа желаемого размера *k*. Однако известно, что этот метод  дает смещенные оценки частот встречаемости мотивов, что демонстрируется на данном слайде. Считаются математические ожидания выбора при случайном блуждании треугольника в разных графах, имеющих одинаковое значение пути длины два ребра. Как видите, математические ожидания не равны друг другу, а значит статистическая оценка смещенная.



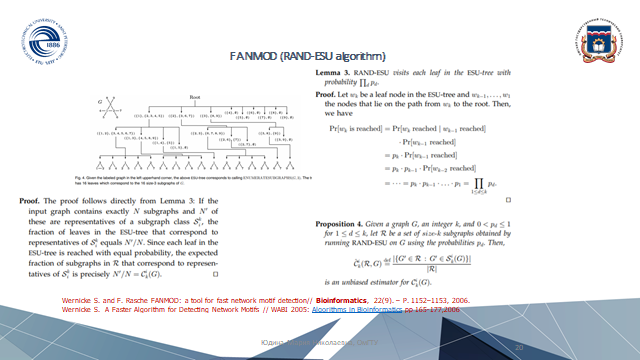
# Методы перебора с отсечением выборки

Среди известных стат. методов, можно выделить метод случайного выбора ребра, но он дает смещенные оценки частот встречаемости мотивов. Более популярным является метод, который мы называем методом Верника – это метод посл. перебора всех мотивов с возможностью усечения генеральной совокупности. Мотивы перебираются путем построения дерева мотивов, на заданном ярусе которого находятся все сетевые мотивы, имеющие соответствующий размер: вершины, ребра, 3-мотивы, можно расширить до 4-мотивов и т.д. Вообще, в статьях Верника указывается, что мы можем отсекать с одинаковой вер-ю ветви на одном уровне, получая насмешенную оценку (аргументация следующая: вер-ть, что мотив будет оставлен в каждом случае можно выразить через произв. вер-тей, что не будет отсечений ни на одном из ярусов). Таким образом данный метод основан на переборе всех подграфов и на построении корневого дерева, листьями которого являются связные подграфы заданного размера. Каждому *k*-му уровню дерева ставится в соответствие некоторая вероятность *pk* сохранения ветвей.



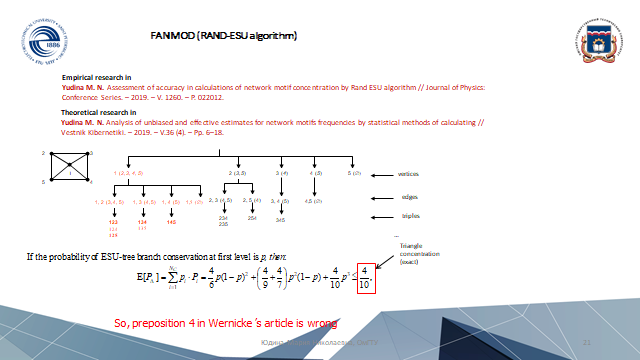
# FANMOD (RAND-ESU algorithm)

 Но действительно ли получаемой при этом оценка несмещенная. С одной стороны в статье Верника, доказывается, что получаемая оценка – несмещенная. А сам алгоритм реализован в программе FANMOD и в igraph.



# FANMOD (RAND-ESU algorithm)

Впрочем, как ни странно, но это не так. Рассмотрим граф и соответствующее ему ESU-дерево, у нас имеется три яруса, на первом ярусе вершины, на втором, все ребра, на третьем – все тройки вершин. Сразу заметим, что если мы отсекаем вершину 1 на первом ярусе, то мы отсекаем большую часть сетевых мотивов (7 из 10) и все треугольники. А если мы отсекаем вершину 5 – мы не потеряем ни одного из мотивов. Очевидно, сколько мы получим в итоге сетевых мотивов, вообще говоря, зависит от сделанного выбора на первом шаге. Это зависимые величины, случайные отсечения на более ранних ярусах влияют на возможности осуществления отсечений на последующих.



# Ошибки в исследованиях

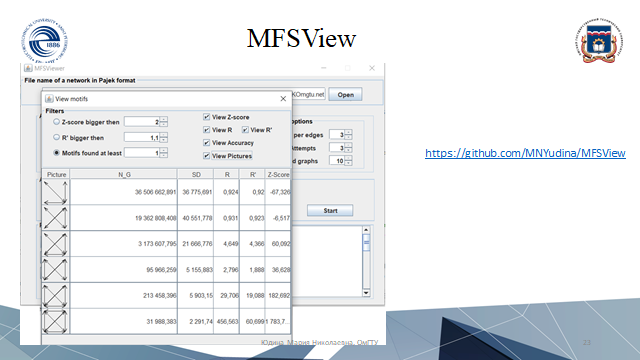
Вообще-то, в современной науке стало нормально, что в половине статей люди предлагают гипотезы, которые они называют теории, <клик>а в половине эти гипотезы опровергаются, как это сделал сам Верник, виртуозно разделавшись со статьей  Роя Ицхека, в которой тот предложил достаточно неэффективный алгоритм и смел критиковать алгоритм самого Верника<клик> Но хуже всего – это не теории, а недомолвки.

Вот вам мой пример – это программа AccMotif для расчёта сетевых мотивов полныйм перебором. Она замечательно и быстро работает. Она скачивается и радует пользователей. Но вот незадача, авторы в ней не пишут, что если загрузить граф чуть больше 20 000 вершин все зависнет навсегда, что память расходуется экспоненциально. А я почти поверила, что это лучшая программа.



# MFSView

Как считать? Пользуйтесь моей программой MFSView. Я постараюсь далее кратко изложить суть этого алгоритма.



# Метод Монте-Карло

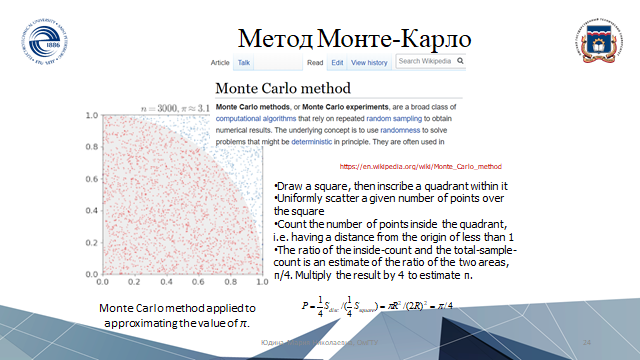
Поскольку мой метод относится к т.н. классическим методам методам Монте-Карло. Поэтому я начну с  объяснения метода Монте-Карло.

Причем, начну с примера. <Клик> Пусть нам нужно рассчитать число ПИ. Чтобы это сделать методом Монте-Карло будем в квадрат единичной площади равновероятно выбрасывать случайные точки. При этом будем считать статистическую оценку доли точек, которые оказываются от начала координат на расстоянии не больше единицы (эти точки красные). Зачем мы считаем эту долю? Потому, что она из геометрических соображений должна быть равно Пи/4. Вот эти соображения.  <Клик> Отношение круга с радиусом R к площади квадрата со стороной 2R.

Итак, при большом числе опытов доля точек изображенным красным сходится к Пи/4, умножая на 4, получим оценку пи.

Выбрасывая все большее число точек, мы получаем все более точную оценку

Короче, к методам Монте-Карло относятся такие статистические методы, в рамках которых используется некоторый случайный процесс,  при многократной реализации этого процесса рассчитываются некоторые эмпирические оценки, используя которые мы можем получить искомые характеристики какой-то другой математической модели

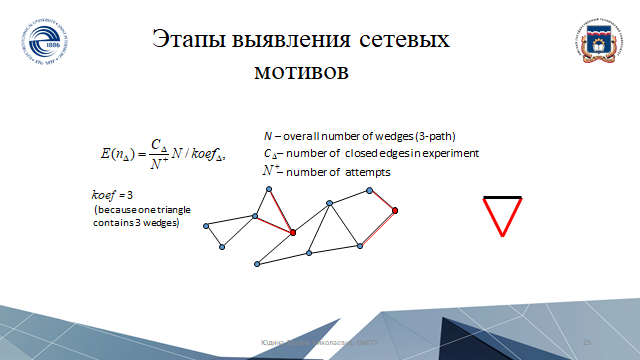


# Случайный выбор триплета

Итак, методы случайного выбора пути – это экземпляры методов Монте-Карло. Рассмотрим случайный выбор пути для оценки числа частот подграфов на трех вершинах. Подграфы на трех вершинах – это либо незамкнутый путь на двух ребрах, либо замкнутый (т.е. треугольник). Главное, организовать равновероятный выбор пути длины 2 ребра.

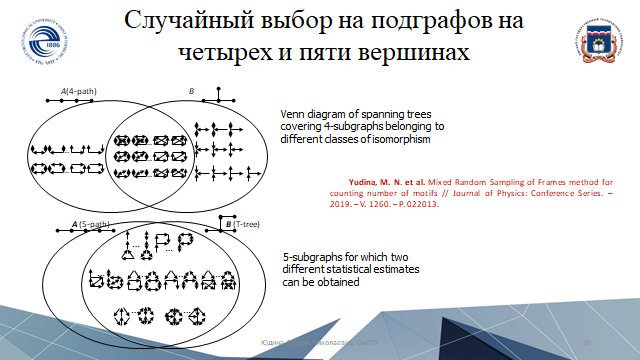
<Клик -5 раз> выбор пути осуществляется, через выбор вершины v c вероятностью pv, пропорциональной числу путей серединой которой она является.<Клик>

После того как мы научились быстро выбирать путь мы оцениваем доли путей, между свободными концами которых существует ребро. Ну а оттуда по формуле и общее число треугольников.



# Случайный выбор на подграфов на четырех и пяти вершинах

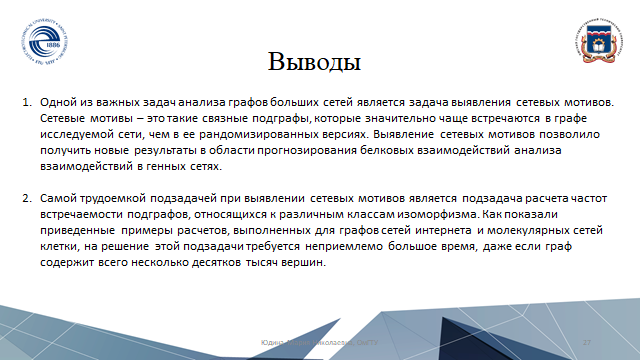
Когда мы предложили метод случайного выбора остовных деревьев, для того чтобы позволить исследовать не только неориентированные подграфы, но и ориентированные и смешанные. Ситуация, когда получаются несколько различных оценок стала возникать чаще: потому, что необходимо проведение статистических экспериментов по всем классам изоморфизма остовных деревьев. Так, например, на четырех вершинах могут быть получены две различные статистические оценки, а на пяти – три, на шести - 6



# Выводы

1. Одной из важных задач анализа графов больших сетей является задача выявления сетевых мотивов. Сетевые мотивы – это такие связные подграфы, которые значительно чаще встречаются в графе исследуемой сети, чем в ее рандомизированных версиях. Выявление сетевых мотивов позволило получить новые результаты в области прогнозирования белковых взаимодействий анализа взаимодействий в генных сетях. Преобладание тех или иных сетевых мотивов является особенностью всякой большой сети, отличающей ее от других сетей, в первую очередь, от сетей другой природы. Так, выявление FULLY CONNECTED TRIAD мотива является признаком графов информационных сетей интернета, для графов молекулярных сетей клетки характерны BI-FAN и FEED-FORWARD LOOP мотивы, в графах трофических сетей обнаруживаются THREE-CHAIN и BI-PARALLEL мотивы и т.д. Тем не менее, подход к исследованию графов сетей на основе сетевых мотивов часто подвергается критике, и как чисто «топологический», и как чисто «статистический» подход.

2. Самой трудоемкой подзадачей при выявлении сетевых мотивов является подзадача расчета частот встречаемости подграфов, относящихся к различным классам изоморфизма. Как показали приведенные примеры расчетов, выполненных для графов сетей интернета и молекулярных сетей клетки, на решение этой подзадачи требуется неприемлемо большое время, даже если граф содержит всего несколько десятков тысяч вершин. При этом современные сети интернета содержат миллионы узлов и связей, а значит, использование только параллельных или распределенных вычислений не позволит обеспечить удовлетворительное решение данной подзадачи. Поэтому необходима разработка статистических методов расчета частот встречаемости подграфов.



# Cytoscape

Особенно нужно быть осторожным с использованием программы Cytoscape. Поскольку для выявления частот встречаемости подграфов эта программа использует один из алгоритмов полного перебора.

