Конспект лекции

Юдин Е.Б., Юдина М.Н.

Введение в моделирование Сложных сетей

# Введение в моделирование Сложных сетей

Здравствуйте, я представляю лекцию 1 “Введение в моделирование сетей и сетевых процессов” 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 разработал Юдин Е.Б. в соавторстве с Юдиной М.Н.



# Сети

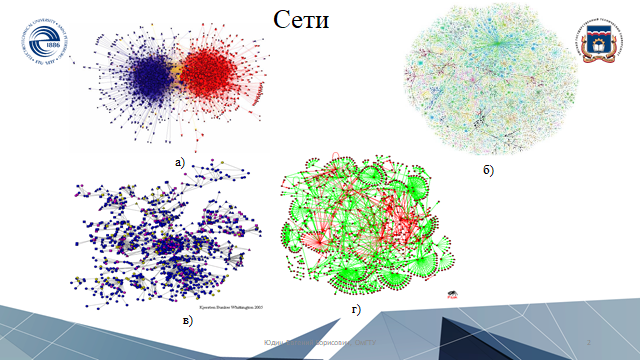
В этой дисциплине мы будем изучать новую отрасль знаний, которая называется Наука о сетях. Она включает в себя исследование сетей телекоммуникаций, таких как сети Интернета, социальные сети, а также т.н. Сетевую медицину (исследование генных сетей, сетей белковых взаимодействий, сигнальных сетей и т.д.).

На данном слайде на рис. а) изображена сеть политических блогов. Раскраска означает политическую поляризацию: синие вершины графа моделируют блоги с преимущественно республиканской приверженностью, красные вершины – с демократической.

На рис. б) представлена графовая модель сети Интернет, вершины – это модели автономных систем (это такие крупные провайдеры типа Ростелекома с внутренней адресацией, таких узлов в мире порядка 60 000), а ребра –  физические соединения между автономными системами.

На рис. в) изображена графовая модель сети сотрудничества ученых в области биотехнологий (исследовалось авторство на патенты с 1976-1998 в области биотехнологий), цветом вершин задается пол патентообладателя, формой вершин – область исследования.

На рис. г) представлена сеть белковых взаимодействий. Современная медицина становится индивидуализированной. Можно по генетическому паспорту определить болезни. Впрочем у нас в стране, насколько мне известно, генетический паспорт декодировать и интерпретировать можно в двух местах



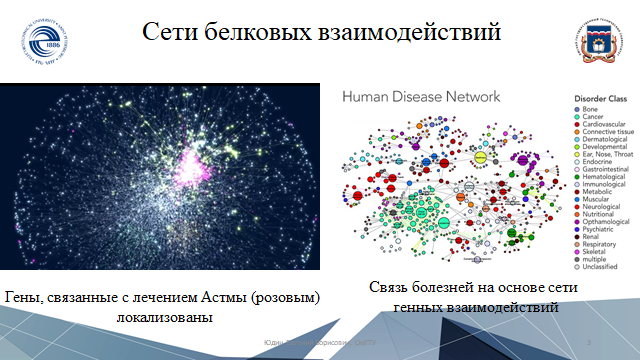
# Сети белковых взаимодействий

На западе исследования генных сетей, сигнальных сетей, сетей белковых взаимодействий активно финансируется, например, такой организацией как NCI (Национальный институт онкологии). Причём теории Сетей используются, чтобы управлять молекулярными механизмами подавления рака. При построении сетей используются совершенно разные подходы: геномная протеомика, метаболомика, а также различные виды анализа медицинских изображений, в том числе гистологические изображения.

Сети оказывают полезными при изучении человеческих болезней в рамках наука о Сетевой Медицине (Network Medicine)

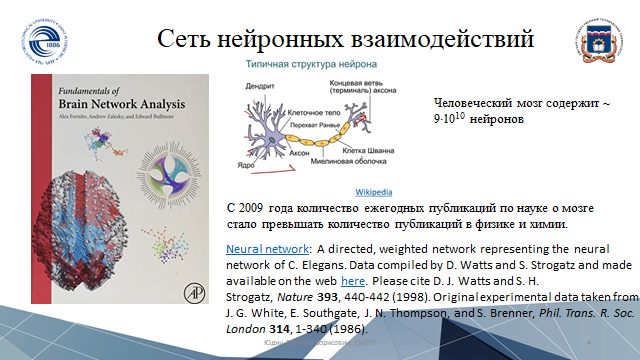
Что это за наука?! Оказывается, что есть генная сеть человека, в которой гены связаны различными взаимодействиями друг с другом. Мы знаем 20-40% этой сети. Но если мы cконцентрируем свои усилия,  то обнаружим что есть подсети, взаимодействия генов некоторых болезней, сходных по симптоматике, очень похожи.

На данном слайде на рис слева  выделены гены ответственные за Астму (а также Хронической обструктивной болезни легких и ряда других схожих болезней), и лекарства, который будет принимать человек, будут влиять на функционирование этих и только их. Это касается не только Астмы, в рамках лабораторных необходимо скачать данные по другим болезням (ишемическая болезнь сердца, разные виды раков). Доступ к сетям вы получите через  систему Cytoscape. Также можно выделить гены, которые проявляются при возникновении нескольких болезней. Так мы можем построить сеть справа, связи болезней между собой.



# Сеть нейронных взаимодействий

Конечно, одной из областей исследования сетей является связи нейронов в мозгу. Но в плане наших знаний о мозге человека, здесь ситуация много хуже ситуации с сетями белковых взаимодействий. Самих нейронов существенно больше генов, а число связей между нейронами – астрономическое число. Впрочем, и число нейронов в некотором смысле астрономическое число. В нашем мозге содержится около 100 миллиардов нейронов — это больше, чем количество звезд в Млечном Пути. Впрочем, это всего лишь 10% от всех клеток мозга, которые у нас есть. Но нейросети некоторых организмов почти полностью изучены. В частности, на данном слайде приведена ссылка на структуру нейросети червя C. Elegans, 302 нейрона и чуть более 5000 синапсов.



# Социальные сети и веб

На данном слайде представлены т.н. эго-сети социальной сети “Вконтакте”. Т.е. берем узел, например одного из преподавателей ОмГТУ(слева) и студента (справа) и смотрим его друзей, а также выявляем отношения уже имеющимися узлами. Что видно. Наука о данных вещь часто субъективно, мы часто видим что хотим. Что вижу я? Вот студент. Вот кластер его одногруппников, вот кластер его бывших одноклассников. И ряд знакомых. Не знаю, два из этих узлов – точно преподаватели. Мне кажется, такая структура означает отсутствие жизни вне университета или школы. Но тут возникает другой вопрос: почему преподаватели (а один из них я) никак не связан с другими студентами. Говорят, современное поколения с детства ведут аккаунты для будущего работодателя, но это вам виднее. Для меня здесь много загадок. Но интересно.



# Социальные сети и веб

На данном слайде слева представлена сеть пользователей вконтакте, указавших местом учебы ОмГТУ. Удивительно, что большинство вошло в один компонент связности. Кто здесь самый влиятельный? У кого больше всего связей? Кто удален от других наименьшим числом связей. Ответ на это вопрос можно получить с помощь различных библиотек и визуальных систем анализа типа Gephi Pajek и т.д. Об этом мы сегодня поговорим.

Справа представлена  эко-сеть вокруг omgtu.ru, на самом деле были собраны данные о структуре веб-страниц ведущих университетах Омска и получены интересные данные, откуда много ссылок, как и куда ссылаются. Но построение эго сети для выбранного узла, было использовано и для изменения структуры веб-ссылок одной образовательной организации, не ОмГТУ, когда представители увидели какая структура у них и у других, они ужаснулись и решили что-то изменить.



# Сети в экономике

На данном слайде слева представлена сеть «мировой закулисы», состоящей из связанных деловыми связями 1318 компаний, контролирующих 60% реального сектора экономики (насколько я помню те, уже давние исследования менее 150 компаний контролировали  40%).

Справа представлена т.н. сеть рекомендаций фильмов. У нее есть особенности. Фильмы между собой не связаны, пользователи то же. Сеть представляется двудольным т.н. графом (есть два компонента, элементы которых связаны между собой). Из этого двудольного графа можно получить различные представления.



# Сети ключевых слов

Как можно видеть на слайде, сети используются также при анализе ключевых слов в научных публикациях. Тем самым можно выделить наиболее горячие тематики исследований.



# Наукометрические сети

Другим примером являются сети подобия конференций, сети соавторства ученых (факт соавторства обозначается связью между узлами)



# Определения

Языком описания всех вышеперечисленных систем является теория графов. Введем некоторые определения.

Математически сеть представляет собой граф G, который определяется через множество вершин V и множество ребер E между этими вершинами:

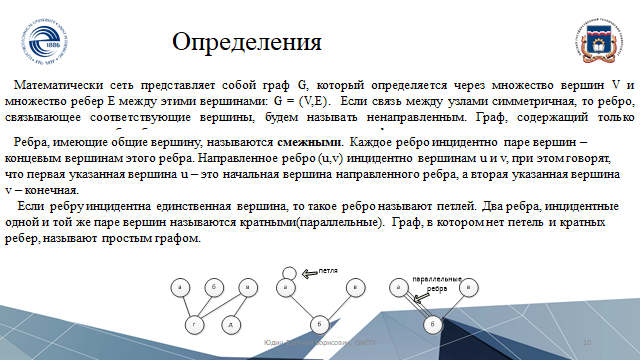
G = (V,E).

Если связь между узлами симметричная, то ребро, связывающее соответствующие вершины, будем называть ненаправленным. Граф, содержащий только ненаправленные ребра, будем называть неориентированным графом. Если связь несимметричная, то ребро будет направленным. Граф, в котором некоторые ребра могут быть направленными, а некоторые ненаправленными будем называть смешанным графом.

Ребра, имеющие общие вершину, называются смежными. Каждое ребро инцидентно паре вершин – концевым вершинам этого ребра. Направленное ребро (u,v) инцидентно вершинам u и v, при этом говорят, что первая указанная вершина u – это начальная вершина направленного ребра, а вторая указанная вершина v – конечная.

Если ребру инцидентна единственная вершина, то такое ребро называется петлей. Два ребра, инцидентные одной и той же паре вершин называются кратными(параллельные). Граф, в котором нет петель и кратных ребер, называют простым графом.

В дальнейшем будем рассматривать только простые графы. Это наиболее распространенный вид графов, подходящей для описания большинства сетей



# Общие характеристики

Рассмотрим некоторые характеристики графа в целом

Средняя степень связности <*k*> вершин – характеристика, определяющая среднее арифметическое степеней связности deg(*vi*) вершин *vi* графа G(V,E): *vi* – это очередная i-я вершина, от одного до некоторого числа задающее число вершин в графе, т.е. от 1 до норма V

,

где deg(*vi* ) – это число соседних вершин для вершин *vi*. Знак сигма – означает, что будет суммирование по всем вершинам

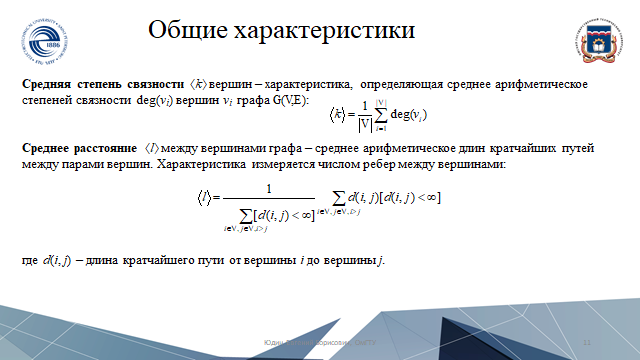
А после того, как по всем вершинам мы просуммируем число соседей для каждой вершины, полученное значение мы должны поделить на общее число вершин (т.е. норма V).  Таким образом, получаем среднее арифметическое степеней связности deg(*vi*) вершин

Среднее расстояние <*l*> между вершинами графа – среднее арифметическое длин кратчайших путей между парами вершин. Характеристика измеряется числом ребер между вершинами:



Последняя формула считается так. Берется сумма *d*(*i*, *j*) – т.е. сумма длин кратчайших путей от всех вершины i до вершины j, конечно, если такой путь существует (ведь он может и не существовать). Если нет пути от *i* до *j* то слагаемое равно нулю. Вот это скобка d(*i*, *j*) меньше бесконечности будет равна нулю (а иначе она будет равна 1) Это так называемая нотация Айверсона.

Хорошо мы сумму путей от каждой вершины до каждой посчитали. Чтобы получить среднее арифметическое, нужно ещё поделить на количество таких путей. Число таких путей нам даёт знаменатель, сумма всех путей это будет сумма выражений в форме Нотаций Айверсона, задающая единицу, если между соответствующей парой есть путь, и ноль - иначе.



# Общие характеристики

Следующий важный показатель - это Коэффициент кластеризации C’ рассчитывается как отношение утроенного  числа   «треугольников» в графе к числу  «вилок», т.е. числу путей длиной два.



Вилка – это путь длиной два ребра. Треугольник – это замкнутый путь на трех ребрах.

На слайде приведен небольшой пример. Изображен граф, для которого число треугольников равно 1 (вы видите, что треугольник один), а число «вилок»  равно 5(4-2-1, 4-2-3, 1-2-3, 3-1-2, 2-3-1), глобальный коэффициент кластеризации по формуле: C ' = 3×1/5 = 0,6.

Коэффициент кластеризации – степень, определяющая насколько узлы стремятся к кластеризации. Например, в сети друзей это вероятность того, что 2 моих друга являются друзьями между собой. То есть это некоторая оценка фрагментированности сети. При высокой кластеризации можно ожидать, что вирус будет распространяться лишь в определенной подгруппе (кластере). При низкой кластеризации высока вероятность быстрого распространения вируса по всей сети



# Меры важности вершин: степень связности

Есть характеристики отдельных элементов графа. Например, очень популярно понятие мер важности отдельных вершин.

В английской литературе это «важность» называется центральность (centrality) вершин (в русском языке также Москва – центр страны, а кремль – центр Москвы).

Итак, концепция первая. Узел тем важнее, чем больше у него связей – это важность по степени связности

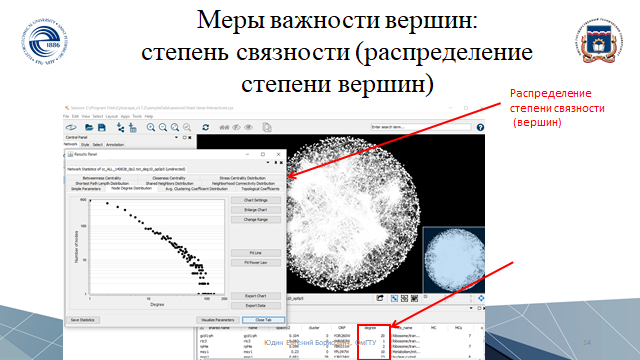
На данном слайде представлен графа, моделирующего сеть семейных кланов Флоренции в XIII-XIV вв.  У семейства Медичи – 6 браков с другими семьями Флоренции, у Строци и Гуадани по 4). В общем случае, чем более активен узел, тем больше у него связей.



# Меры важности вершин: степень связности (распределение степени вершин)

Если в системе Cytoscape мы откроем одну из базовых сеть, например, как на картинке сеть белковых взаимодействий дрожжей, мы увидим степени связности, характерные для каждого белка.

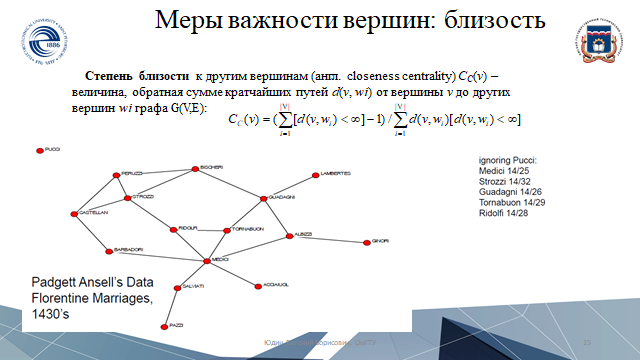
Cytoscape -  биоинформатическая платформа с открытым исходным кодом, которую мы будем изучать. Она предназначена для визуализации сетей молекулярных взаимодействий и биологических путей с возможностью использования дополнительных данных, таких как функциональная аннотация, информация об уровне экспрессии генов и прочих.

Также в Cytoscape можно посмотреть распределение степени связности вершин. По оси абсцисс отложена степень связности, а по оси ординат – число вершин с заданной степенью связности. 

# Меры важности вершин: близость

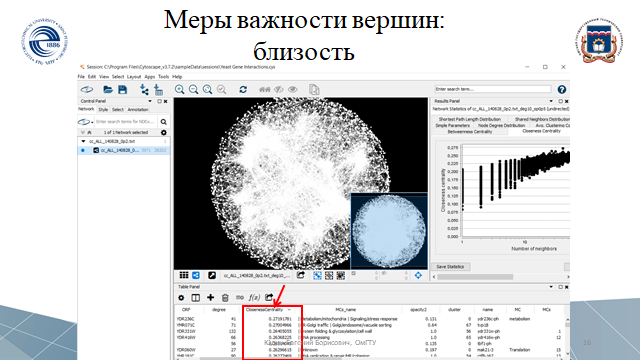
Степень близости – другой показатель насколько узел удален от других узлов в связной компоненте. Чем больше удален (вот тут в формуле сумма всех расстояний) тем меньше важность по этому показателю. Если узел со всеми другими узлами связан ребром, то числитель и знаменатель совпадут. Значимость будет единица.

В сетях телекоммуникаций этот показатель характеризует время отклика сети.



# Меры важности вершин: близость

При анализе в  Cytoscape сети белковых взаимодействий дрожжей, мы увидим, что наиболее важным по мере близости оказываются белки ymr071с и ydr236с  (выделены красным прямоугольником)  Эти белки важны как в сигнальных путях организма и в работе комплекса Гольджи.



# Меры важности вершин: Посредничество

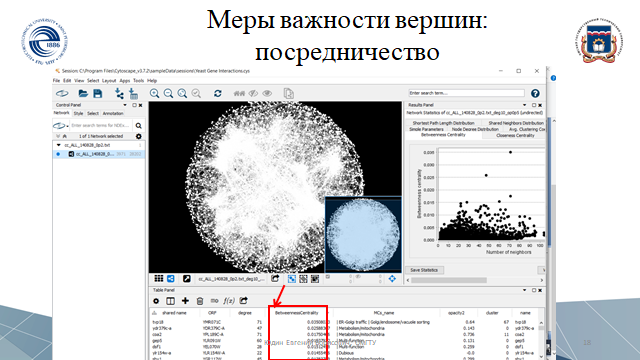
Степень посредничества показывает, насколько часто узел находится на кратчайших путях между другими узлами. Насколько он будет информирован. Чем чаще, тем больше значение. Оно нормировано на выражение, которые некоторые из вас могут знать N(N-1)/2 это число сочетаний из N по два. Между несколькими вершинами могут быть пути. На четырех вершинах таких путей 6 (4\*3 /2).

****

# Меры важности вершин: Посредничество

При анализе в  Cytoscape сети белковых взаимодействий дрожжей, мы увидим, что наиболее важным по мере посредничества оказываются tvp18 - интегральный мембранный белок; локализуется в поздних пузырьках Гольджи

И ydr379C-A -Митохондриальный белок, участвующий в сборке сукцинатдегидрогеназы



# А теперь немного посчитаем

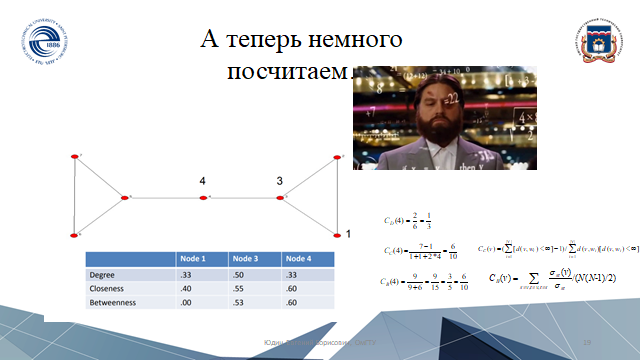
Предлагается посчитать.

Посчитать для графа характеристики важности вершины 4.

По степени связности у четвертой вершины степень связности будет 2. Нормированная характеристика будет равна 2/6 (у вершины 6 возможных соседей) т.е. 1/3.

Для важности по близости у четверки также 6 соседей. По представленным формулам это идёт в числитель. А сумма всех путей до других вершин следующая для 5 и 3 по единичке, и 4 вершины удалены двумя ребрами. Вот и ответ 0,6

Для важности по промежуточности нужно посмотреть сколько кратчайших путей проходит через вершину. Итак из трех вершин слева, чтобы пройти в три вершины справа будет 3 на 3 = 9 кратчайших путей. А сколько всего кратчайших путей 6 на 5 и делить на 2. Будет 15. вот и ответ.



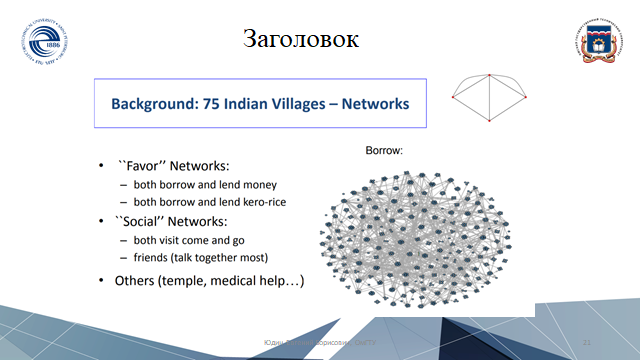
# Использование мер важности по материалам Matthew O. Jackson

Рассмотрим пример применения мер важности. Один из примеров, возьмем из курса Social and Economic Network Analysis, в котором Matthew O. Jackson  рассказал о своём исследовании, проведенном в одном районов в Индии – Карнатака. Это район в основном с сельским населением



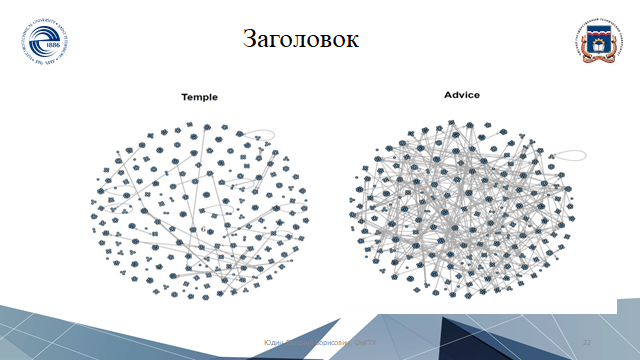
# Объект исследования

Они исследовали 75 деревень в одном районе страны и смотрели на сети взаимодействий. Кто у кого занимает деньги. Вы видите эту сеть на слайде



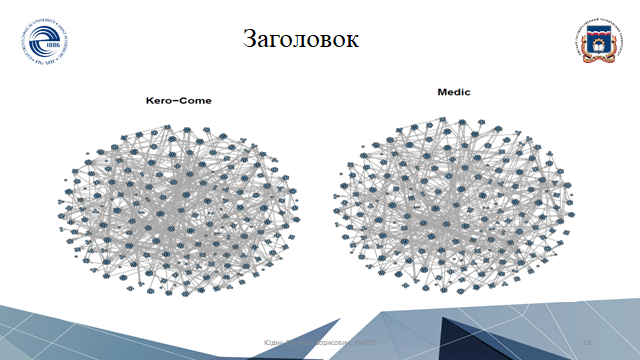
# Религиозная сеть и сеть взаимодействия семей

Кто ходит в чей храм, кто кому дает жизненные советы. Вы видите эти сети на слайде.



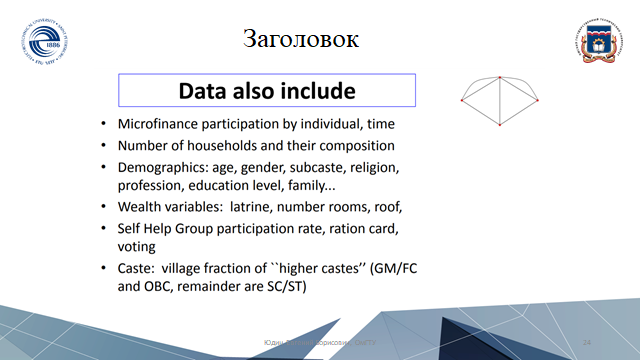
# Керосин и медицина

Кто у кого покупает керосин, кто у кого лечится. Похоже, по исследования Мэтью О Джексона, лечатся они у друг друга.



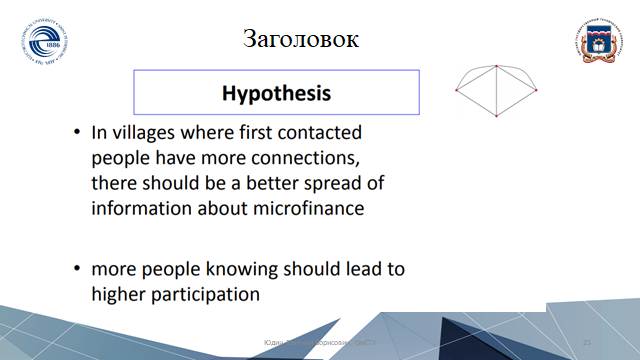
# Использование мер важности по материалам Matthew O. Jackson

Зачем проводилось это исследование. Ну не просто же так. Конечно нет. Для того чтобы узнать, кого сделать ответственным за продвижение микрофинансовых организаций в деревне.



# Гипотеза

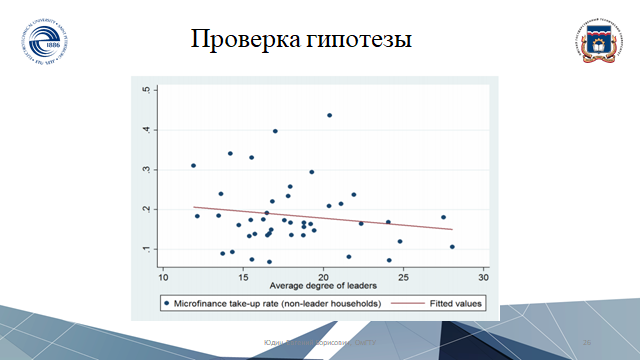
Какая была гипотеза. 1) У кого больше друзей, тот и лучше сможет распространить влияние микрофинансирования в деревне.



# Проверка гипотезы

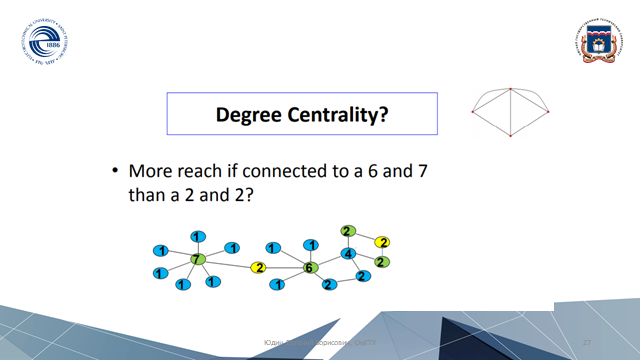
И каков результат? Посмотрим график. В каждой деревне был один человек. И он отмечен точкой. Имеем 75 точек. По оси абсцисс число связей лидера, по оси ординат его влияние со временем на продвижение микрофинансирование. И как сработала гипотеза?

Если бы она была правильная, то чем больше степень связности, тем должно быть больше продвижение. Т.е если строить линейную регрессию, то  она должна идти на подъем. А мы видим, что она напротив – скорее идет вниз.



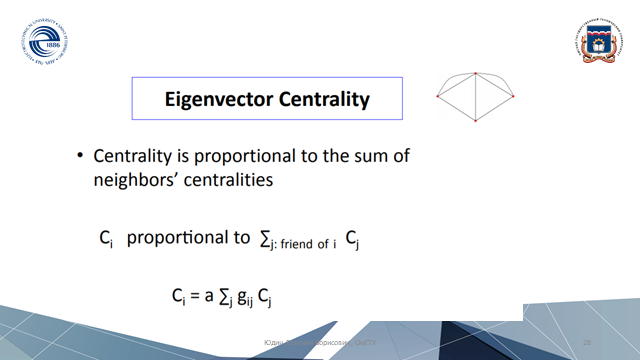
# Почему гипотеза не работает?

Если объяснять почему так может происходить, нужно посмотреть на этот рисунок. Какая двоечка важнее? По центру или с краю. Наверное центральная. Она связана с вершинами у которых и так связей много.  Может в этом причина?



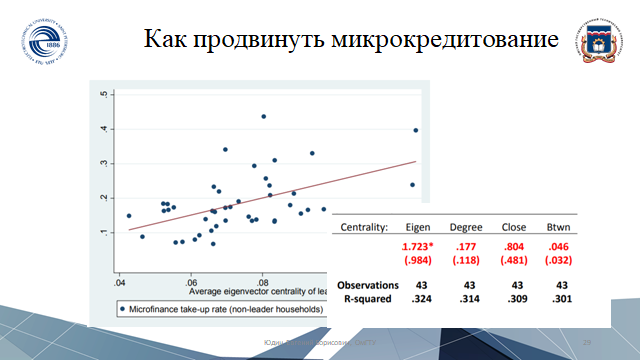
# Важность по значений собственного вектора

Вообще говоря, существует ряд других мер важности, которые мы сегодня не успеем пройти. Одна из известных это т.н. важность на основе значения собственного вектора. Важность вершины зависит от важности соседних вершин. Это похоже на алгоритм в гугл, когда важность сайта определяется от сайтами, на которые много ссылаются. Это задача решается на основе решения задачи нахождения собственного вектора. Многие пакеты эти могут рассчитать для графов разного размера.



# Как продвинуть микрокредитование

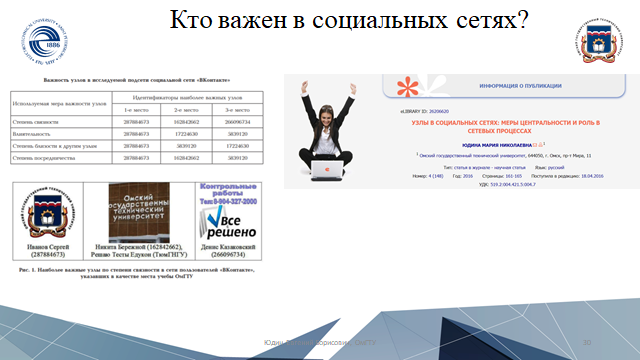
Исходя из проверки, представленной на слайде, мы видим, что именно важность по собственному вектору обеспечивает максимальное продвижении микрофинансирования. А значит в следующий раз в странах третьего мира эти компании смогут эффективнее выбирать адептов для продвижения микрофинансирования в странах третьего мира. Пожелаем американцам успеха в их нелегком деле -) и перейдем на завершающие слайды.



# Как продвинуть микрокредитование

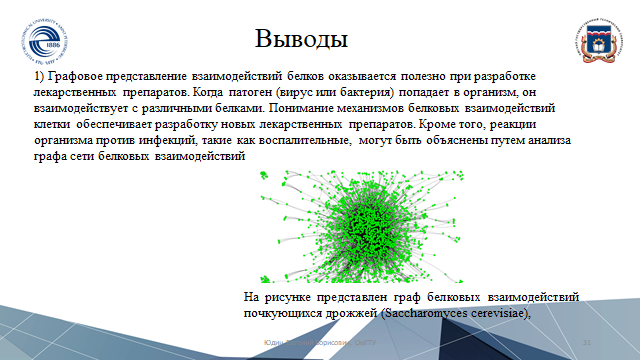
Вас, вероятно, больше интересует, кто важнее с студенческих сетях, правда?

Наши исследования показывает, что это часто официальные лица, фотографы, лица, чей бизнес состоит в помощи студентам



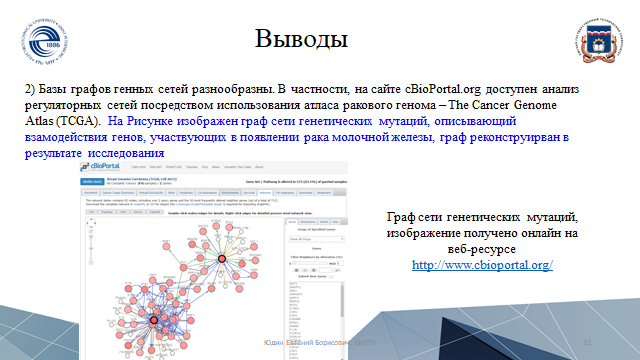
# Выводы

Перейдем к выводам лекции. 1) Графовое представление взаимодействий белков оказывается полезно при разработке лекарственных препаратов. Когда патоген (вирус или бактерия) попадает в организм, он взаимодействует с различными белками. Понимание механизмов белковых взаимодействий клетки обеспечивает разработку новых лекарственных препаратов. Кроме того, реакции организма против инфекций, такие как воспалительные, могут быть объяснены путем анализа графа сети белковых взаимодействий. На рисунке представлен граф белковых взаимодействий Saccharomyces cerevisiae



# Выводы

2) Базы графов генных сетей разнообразны. В частности, на сайте cBioPortal.org доступен анализ регуляторных сетей посредством использования атласа ракового генома – The Cancer Genome Atlas (TCGA). На Рисунке изображен граф сети генетических мутаций, описывающий взаимодействия генов, участвующих в появлении рака молочной железы.



# Выводы

3) Доступны пакеты для анализа молекулярных сетей клетки, одной из популярных программ является программа Cytoscape. Один Одним из способов доступа к исследуемым молекулярным сетям клетки является использование программы CyNDEx 2, которая встроена в систему Cytoscape и позволяет импортировать данные с онлайн-платформы NDEx (Network Data Exchange).

