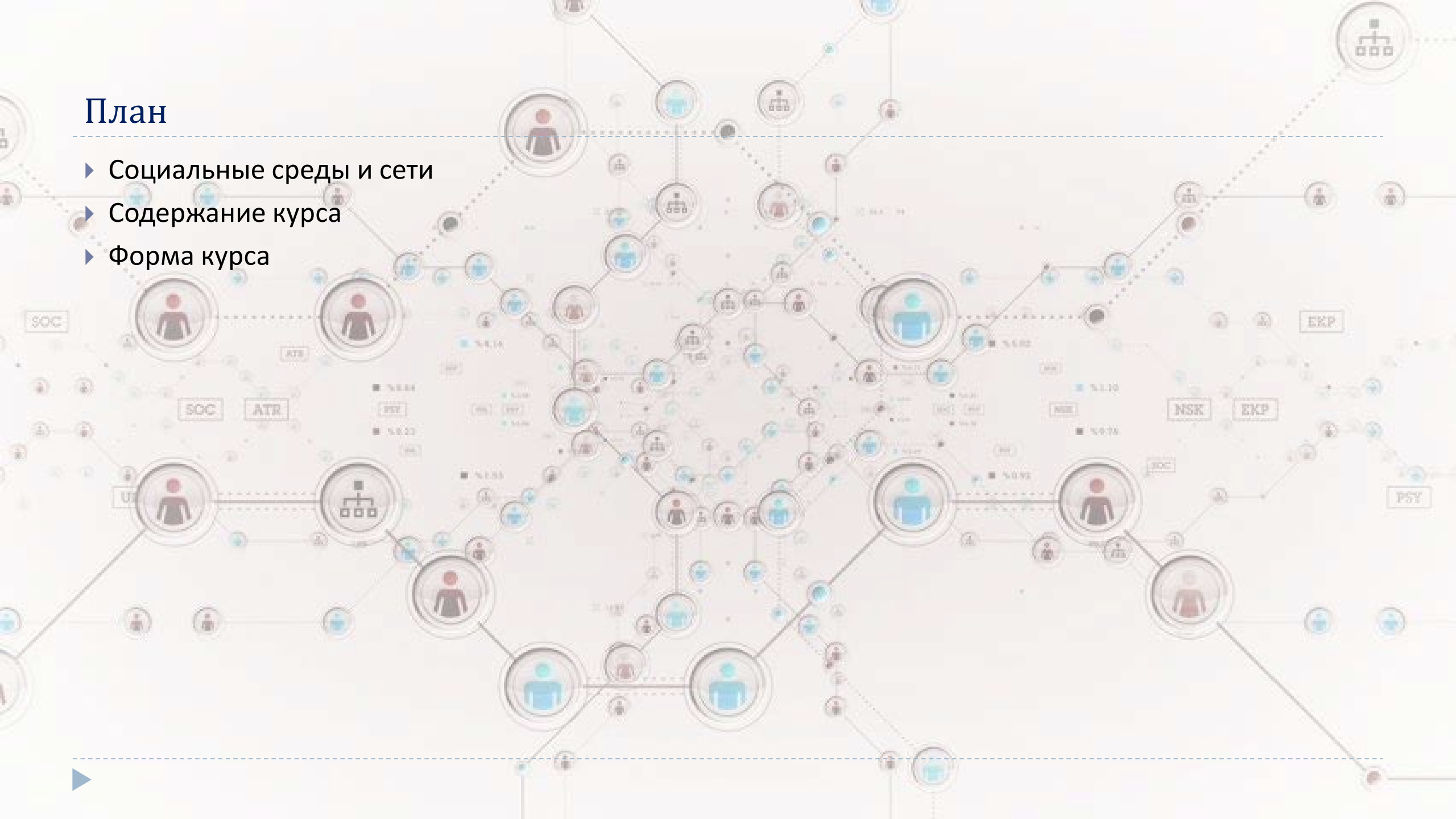
Введение в курс «Анализ Больших данных в социальных средах»

Николай Скворцов
nskv@mail.ru

План

- Социальные среды и сети
- Содержание курса
- Форма курса



Соцсети сегодня

- Распространённые онлайновые социальные сети
  - В России: Facebook, VK, Twitter, LinkedIn, OK
  - В мире также: MySpace, BaiduSpace, Orkut и множество других
- Социальные сети имеют огромное влияние на жизнь людей
  - Доступность социальных связей
  - Влияние на личную жизнь
  - Влияние на политические, экономические, социальные процессы
  - Возможность дешёвого анализа социальных связей и активности
- Однако онлайновые социальные сети – только часть мира социальных сетей и социальных сред

Анализ социальных сетей

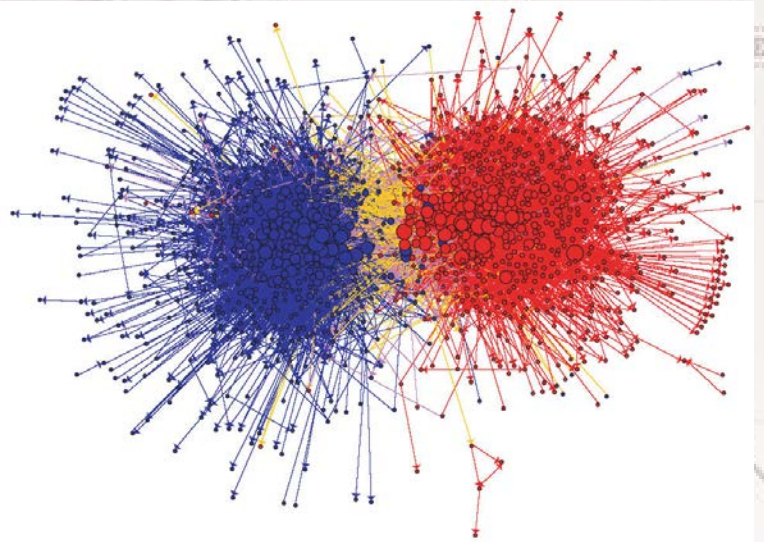
- Возник задолго до распространения онлайновых социальных сетей
- Исследуемые сущности
  - Агенты – субъекты взаимодействия (люди, группы людей, организации)
  - Ресурсы – пассивные объекты в совместном использовании агентами (книги, видео и др.)
  - Связи и взаимодействия - отношения
- Изучает различные свойства определённых типов взаимодействий между агентами с учётом их характеристик

Примеры социальных сетей: Социальные связи

- Виды
  - Родственные
  - Взаимодействия людей
  - Знакомства
  - Территориальные связи
  - …
- Слабые и сильные связи
- Постоянство и изменчивость связей
- Социальный капитал
- Эгоцентрические сети

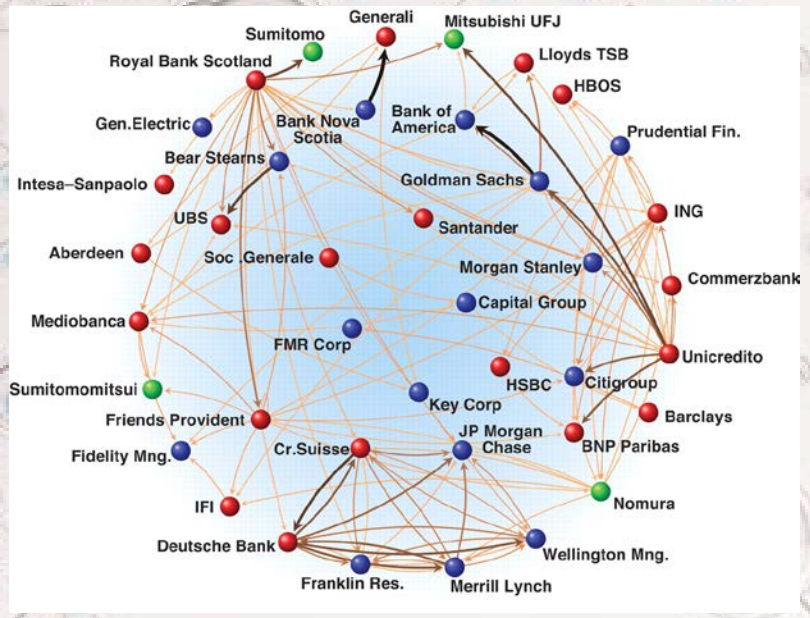
Примеры социальных сетей: Сети в социологии и политологии

- Социология дала большой толчок развитию теории сетей
  - Тесные миры Милгрема
- Подходы к исследованиям
  - Наблюдение - пассивная регистрация взаимодействий агентов (в том числе с датчиков)
  - Эксперименты – активное воздействие на социальную среду
  - Опросы – регистрация мнений о взаимодействии
- Задачи
  - Исследование сообществ
  - Распространение идей
  - Общественное мнение



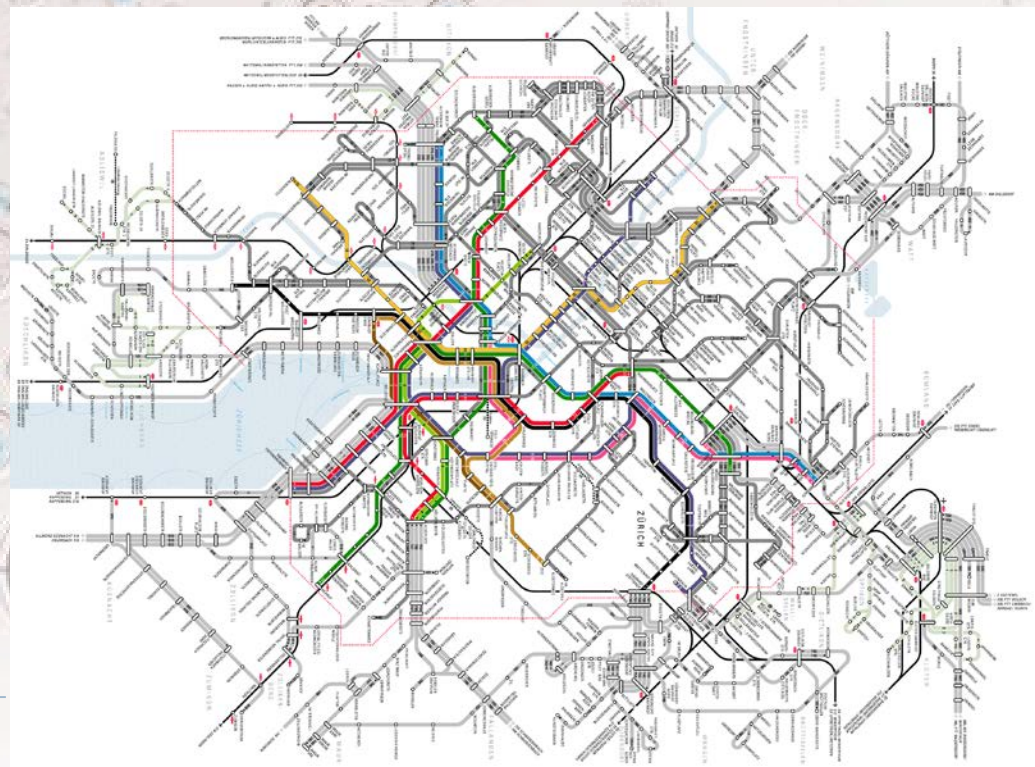
Примеры социальных сетей: экономические взаимодействия

- Виды
  - Предпочтения покупателей
  - Взаимодействие предприятий
  - Финансовая деятельность
- Задачи
  - Рекомендательные системы
  - Потоки товаров и средств
  - Независимые портфели
  - Наилучшие стратегии
  - Модели конфликтов
  - Выявление мошенничества



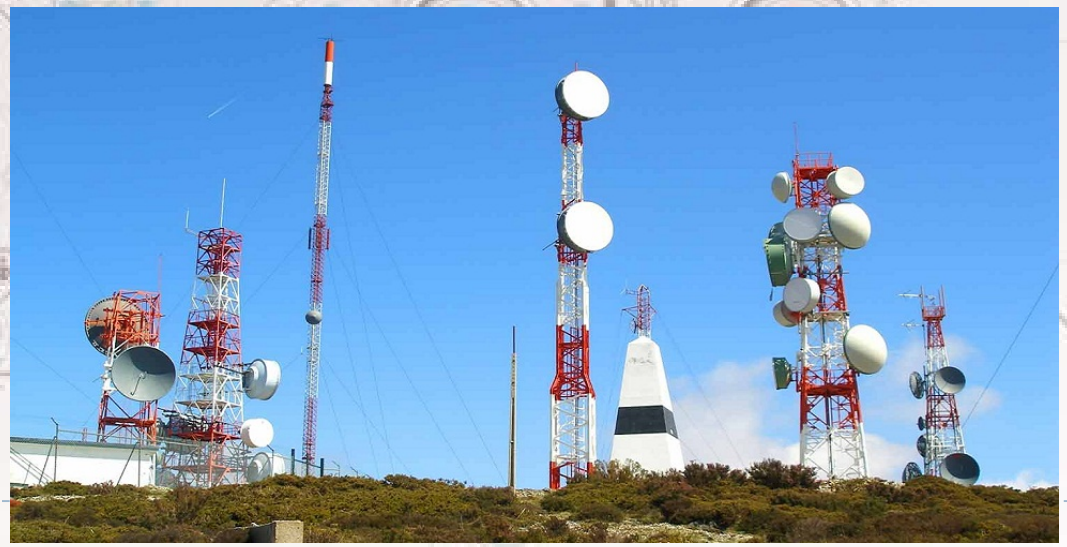
Примеры социальных сетей: транспортные сети

- Виды
  - Транспортные потоки
  - Транспортные системы
- Задачи
  - Связанность
  - Кратчайший путь
  - Длина пути
  - Пропускная способность
  - Мосты
  - Посредничество узлов



Примеры социальных сетей: телекоммуникации

- Виды
  - Звонки
  - Адресные книги
  - Электронная почта
- Задачи
  - Связи пользователей
  - Нагрузка на сеть
  - Обеспечение связи
  - Информационные потоки

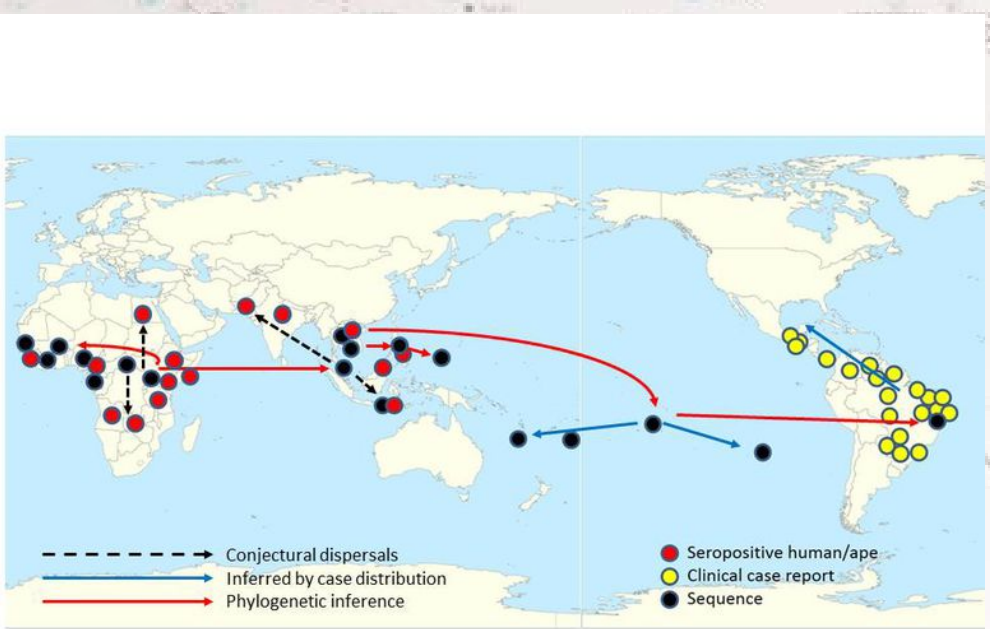


Информационные сети

- Виды
  - Медиа
  - Цитирование
  - Новшества
- Задачи
  - Индексы цитирования
  - Распространение информации
  - Управление
  - Тренды и изменения

Примеры социальных сетей: эпидемиология

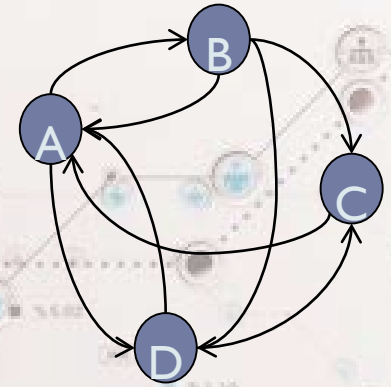
- Задачи
  - Прогноз распространения вирусов
  - Стратегия противодействия распространению
  - Факторы риска заражения
  - Источник заражения



Состав курса

Представление сетей

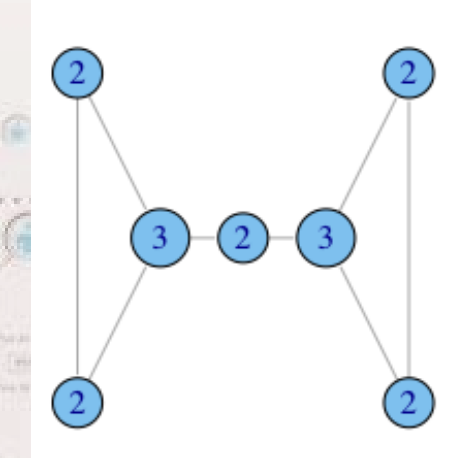
- Графы
  - Вершины – агенты и ресурсы
  - Рёбра или дуги – связи и взаимодействия
  - Атрибуты – свойства агентов и особенности взаимодействий
- Матрицы смежности
- Матрицы двудольных сетей



	A	B	C	D
A	0	1	0	1
B	1	0	1	1
C	1	0	0	0
D	1	0	1	0

Сетевые метрики

- Основные метрики графов
  - Степень, плотность, коэффициент кластеризации, связность, пути и расстояния, диаметр сети, пропускная способность и др.
  - Распределение степеней сети. Средняя длина пути и коэффициент кластеризации
  - Метрики центральности вершин, хабы
  - Особенности для транзитивных, взаимных или ориентированных связей, престиж и влияние вершин
  - Инструменты анализа сетевых метрик

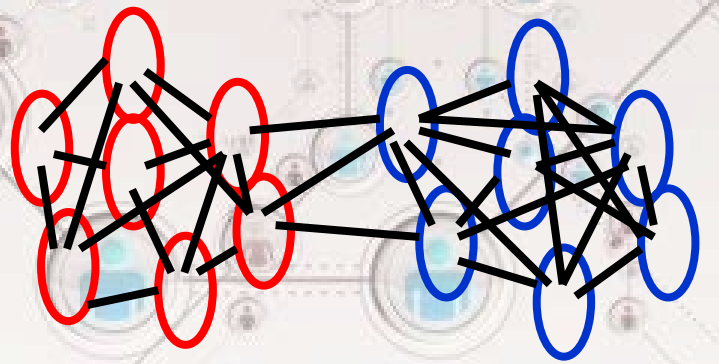


Сетевые структуры

- Вершины разных типов
  - Сходство и эквивалентность вершин
  - Ассортативное смешивание
  - Партнёрские сети и двудольные сети
  - Эгоцентрические сети
- Различные типы связей
  - Роли и положения
  - Взвешенные отношения
  - Позитивные и негативные связи, баланс отношений
- Связность и компоненты сети
  - Достижимость вершин
  - Модульность сети
  - Мосты
- Мотивы – структуры реальных сетей, появляющиеся заметно чаще, чем в случайных сетях
  - Диады и триады

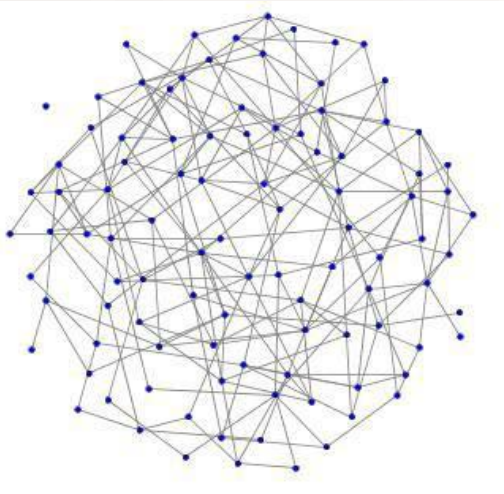
Сетевые сообщества

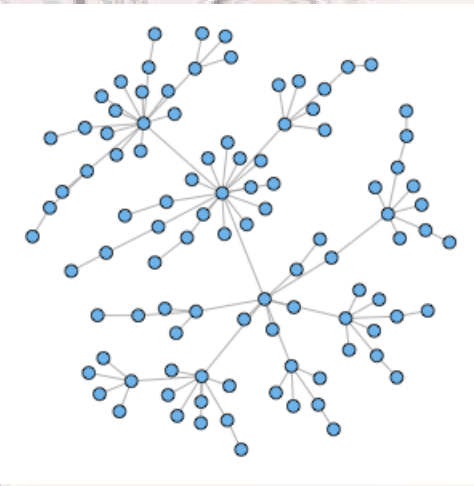
- Различные типы клик
- Кластеризация сети
- Алгоритмы выявления сообществ.
- Перекрывающиеся сообщества
- Сообщества в сетях с ориентированными и взвешенными отношениями.



Моделирование сетей

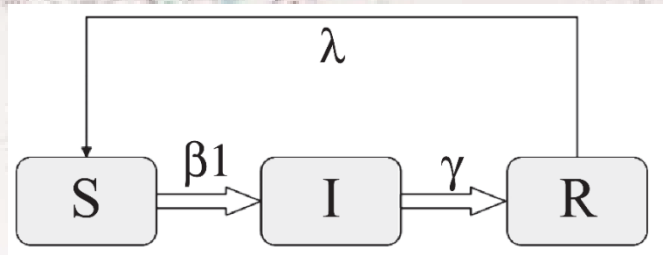
- Модели случайных сетей
- Модели тесных миров
  - Друзья друзей
  - изучение модели, основанной на эксперименте Милгрема
- Законы степенного распределения
  - Изучение моделей, основанных на наблюдении за степенью вершин
  - Модель безмасштабной сети - доля вершин со степенью k пропорциональна k в некоторой степени (от 2 до 3)
  - Модель приближена ко многим реальным сетям
  - Предпочтительное присоединение
- Оценка метрик в моделях сетей
- Модели случайных ориентированных сетей

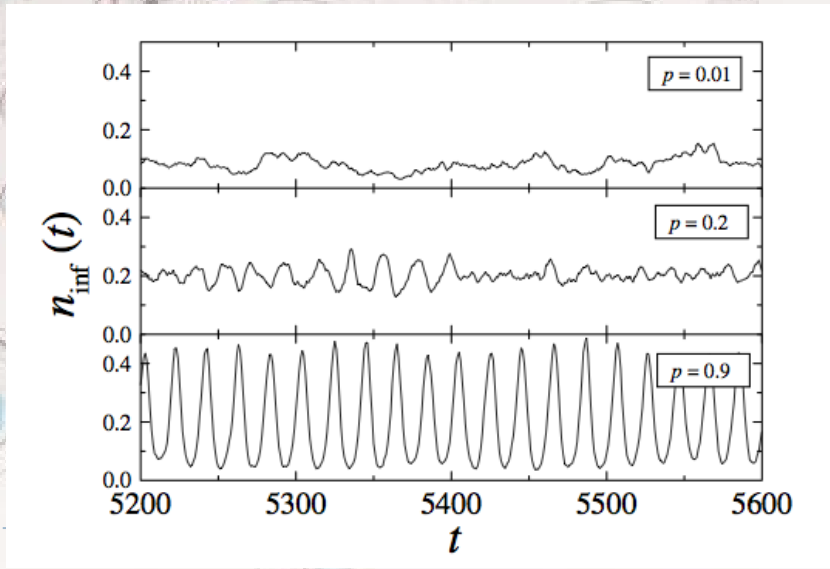




Сетевые процессы

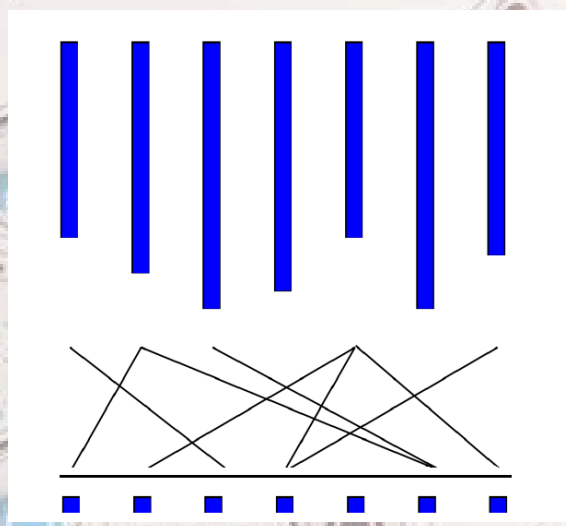
- Проникновение и вирусное распространение в сетях
- Модели распространения
  - SI, SIR, SIS, SIRS
- Каскадное распространение
- Устойчивость сети





Анализ масштабных сетей

- Модели распараллеливания вычислений
  - Bulk Synchronous Parallel (BSP) – параллельное выполнение задач и обмен сообщениями
  - Gather-Apply-Scatter (GAS) – анализ смежных рёбер и посылка сообщений по рёбрам
  - Map/Reduce – принцип Hadoop
- Снижение размерности
  - Выбор влияющих характеристик элементов сети
- Средства распараллеливания алгоритмов над графами



Вопросы доверия данным социальных сетей

- Сложности интерпретации данных в исследованиях
  - Разница между опросом и анализом данных онлайновых сетей
- Психологические аспекты
  - Влияние среды, цели, намерения агента
- Сравнение наблюдаемых взаимодействий и взаимодействий взятых из текста
- Влияние спама, рекламы, платных тем и других воздействий
- Фиктивные агенты
- Подходы к повышению качества данных соцсетей

Доступ к онлайновым соцсетям

- Конкретные социальные сети:
  - Facebook, Twitter, LinkedIn, Google+, VK
- Модели и особенности социальных графов
- API
- Доступ и анализ графов

Другие темы

- Текстовая аналитика в соцсетях
  - Поиск ключевых слов в сети
  - Анализ анкет
- Методы машинного обучения в сети
  - Кластеризация
  - Классификация
- Анализ всплесков, трендов, тенденций
- Методы рекомендательных систем
- И другие

- Связаны ли вершины?
- Насколько они далеки друг от друга?
- Важные вершины (хабы?)
- Сообщества
- Как развиваются структуры, какие процессы их вяжут. => Случайные сети
- Богатый станет богаче => Преобладающее присоединение
- Друзья друзей => Тесный мир
- Оптимизация сетей
- Проникновение информации
- Формирование мнения (консенсус)
- Координация и кооперация
- Устойчивость к атакам (если убрать некоторые вершины, будет ли работать сеть)
- Примеры сетей (ингредиенты и рецепты, предсказание рейтинга рецепта)
- Сеть дельфинов, развитие экономики (страны и их производство, что будут производить в будущем)
- Он-лайн сети (предложение товаров)

Форма проведения курса

- Лекции раз в неделю
  - Четверг 17:00 - 18:30
  - Аудитория 659
  - Кроме 13 октября (конференция DAMDID’2016)
  - Листок присутствия
- Практикум
  - 3-4 темы
  - Показ и возможность самостоятельного использования
- Экзамен
  - По вопросам
  - Учёт посещения

Литература

- S. Wasserman, K. Faust. Social Network Analysis: Methods and Applications. – Cambridge University Press, 1994
- M. E. J. Newman. Networks: An Introduction. –Oxford University Press, 2010
- J. Leskovec, A. Rajaraman, J. D. Ullman. Mining of Massive Datasets. Second Edition. – Cambridge University Press, 2014.
- Social Network Data Analytics. Charu C. Aggarwal (Ed.) – Springer, 2011.