Занятие 1: Введение в анализ социальных сетей и основные показатели

НИУ ВШЭ

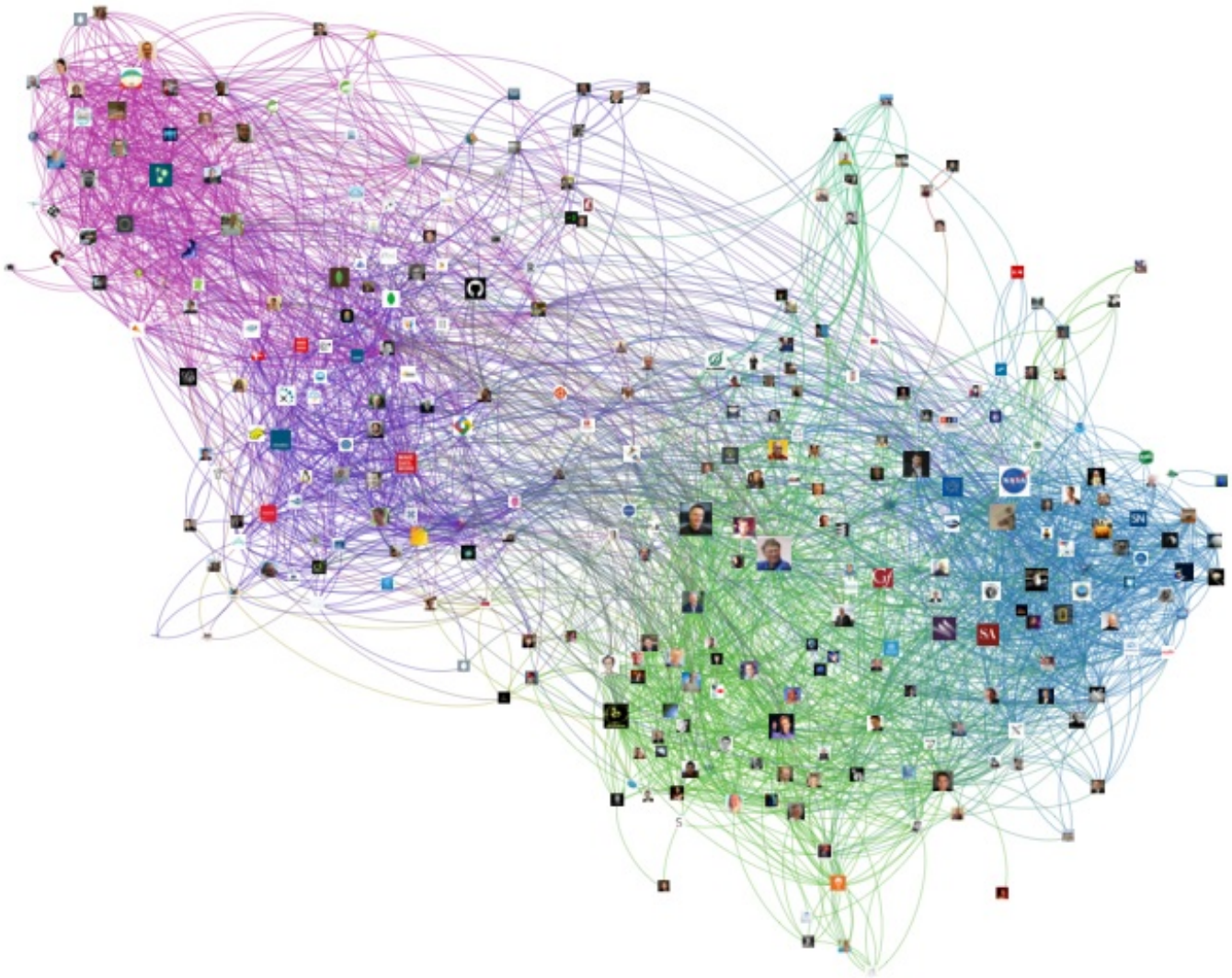
Москва, 2019

План занятий

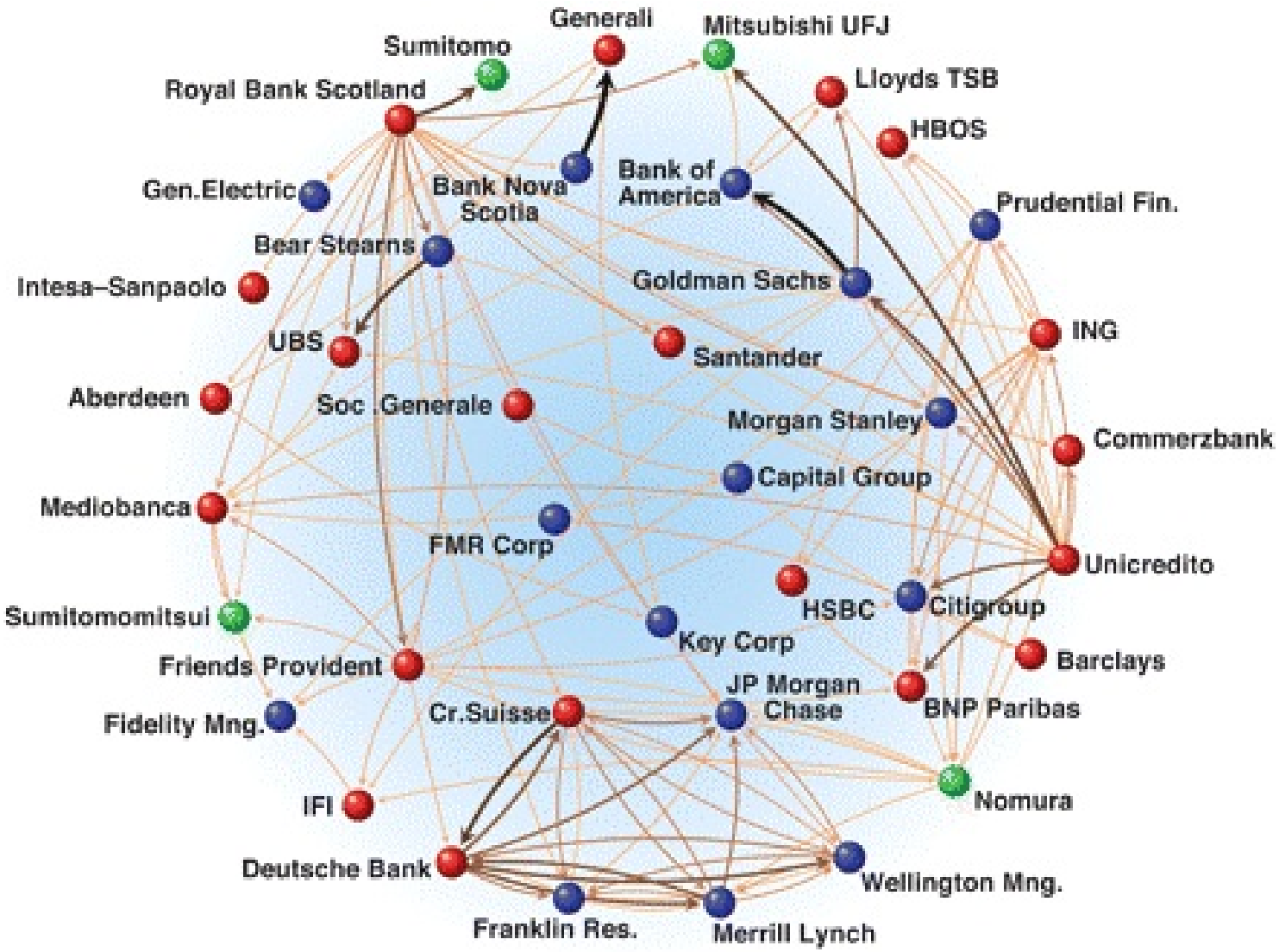
- Введение в сетевой анализ, немного истории сетевого анализа, практическое использование сетей, ключевые описательные статистики сетей, формат работы с сетевыми данными
- Сообщества в социальных сетях
- Модели роста и формирования социальных сетей, Эгосети
- Сбор и загрузка данных из сетей

Социальные сети везде: сеть Facebook

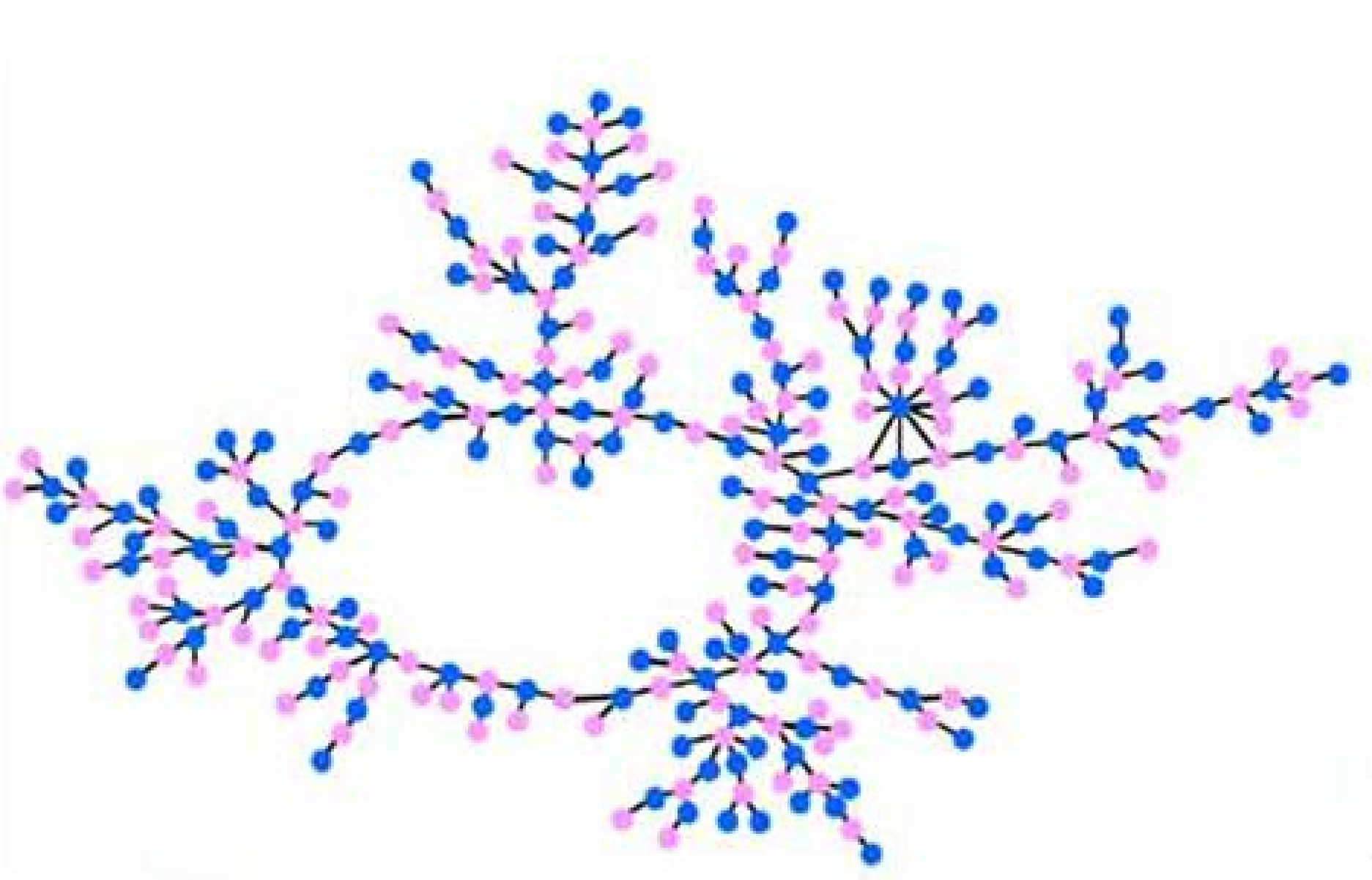
Сеть Twitter



Сеть финансов



Сеть романтических отношений



Сеть аэроперелетов

Что такое сеть?

Сеть (G) - совокупность вершин (V) и ребер (E). G(N, E).
Сеть - network, graph.
Вершина, актор - vertex, node, actor.
Ребро, связь - edge, link, tie.

А зачем?

Чем сетевой анализ принципиально отличается от других видов анализа?
Обычно мы анализируем взаимосвязь между характеристиками пользователя на репрезентативной выборке. Это показывает нам, как что-то взаимосвязано (влияет) на что-то.
Пример: Как взаимосвязана продуктивность сотрудника с его стажем работы в компании/его профессиональным опытом/загруженностью/числом отправленных и полученных писем?
В сетевом анализе фокус смещается на взаимоотношения между людьми, принимая во внимание их индивидуальные характеристики.
Пример: Как взаимосвязана продуктивность сотрудника с продуктивностью его социального окружения? Склонны ли с течением времени сотрудники перенимать продуктивность своего социального окружения?

Сетевой анализ: история возникновения

Анализ социальных сетей базируется на инструментарии теории графов. Теория графов была изобретена Леонардом Эйлером при работе над задачей о кёнигсбергских мостах (1736 год).

Сетевой анализ: история возникновения

- Якоб Морено (1930-е) - психологическое состояние человека зависит от его окружения. Разработал *социометрический подход*.
- Фрэнк Харари - стандартизация инструментария теории графов.
- 1959 г. - Эрдош и Реньи разрабатывают модель случайного графа.
- С 1960-х годов начинается активное развитие направления в общественных науках. Милгрэм проводит эксперимент по изучению реальных социальных сетей.
- С 1990-х годов - активное развитие направления физиками и информатиками.
- 1999 год - появление моделей малого мира (модель Ваттса-Строгатца) и модели предпочтительного присоединения (Барабаши-Альберт).
- После 2000 годов - развитие статистических моделей для анализа сетей (ERGM, SAOM), моделей распространениия информации в сетях и т.д. Особое внимание уделяется анализу больших сетей.

Практическое использование

- HR и People Analytics: изучение структуры взаимодействий между сотрудниками, выявление ключевых лидеров мнений, выявление конфликтов, выявление недоработок в организационной структуре
- SMM: идентификация структуры аудитории, пределение лидеров мнений
- Рекомендации!
- Инвестиции? Изучение структуры инвестиций в стартапы, выявление того, какая структура инвестиций приводит (какие) старпаты к успеху?
- Идентификация недобросовестных клиентов/заемщиков
- Медицина: распространение заболеваний и вакцинация

Как собирают сетевые данные?

- Опросы и интервью (задаем вопросы человеку о его социальном окружении). Возможны смещения из-за социальной желательности.
- Эксперименты
- Загружают из онлайн-источников: соцсети, сайты, систематизированные БД, библиотметрические ресурсы. Возможны смещения из-за дизайна платформ (далее, algorithmic confounding).
- Выводы из текстовых данных: книги, статьи

Как кодируют сетевые данные: форматы данных

- Матрица смежностей (adjacency matrix)
- Список ребер (edgelist)
- Список смежности (adjacency list)
- Graph ML

http://asocialnetworks.blogspot.ru/2018/02/blog-post.html

Какие бывают сети?

- Направленные (ориентированные) и ненаправленные - directed/undirected. Если для всех $x_{ij}=x_{ji}$, то сеть ненаправленная. В противном случае - направленная.
- Взвешенные и невзвешенные. Для невзвешенных сетей x_{ij} может быть 1 или 0. Для взвешенных сетей x_{ij} - все, что угодно, в том числе и отрицательные значения.
- Полные сети - полный граф. Такая сеть, в которой реализованы все возможные связи. Вопрос - сколько связей в полном направленном и ненаправленном графах, если число вершин n?

В чем анализируют сети? Программное обеспечение

Python не идеальная среда для анализа сетей (к сожалению). К еще большему сожалению, идеальной среды для анализа сетевых данных в природе не существует! :(

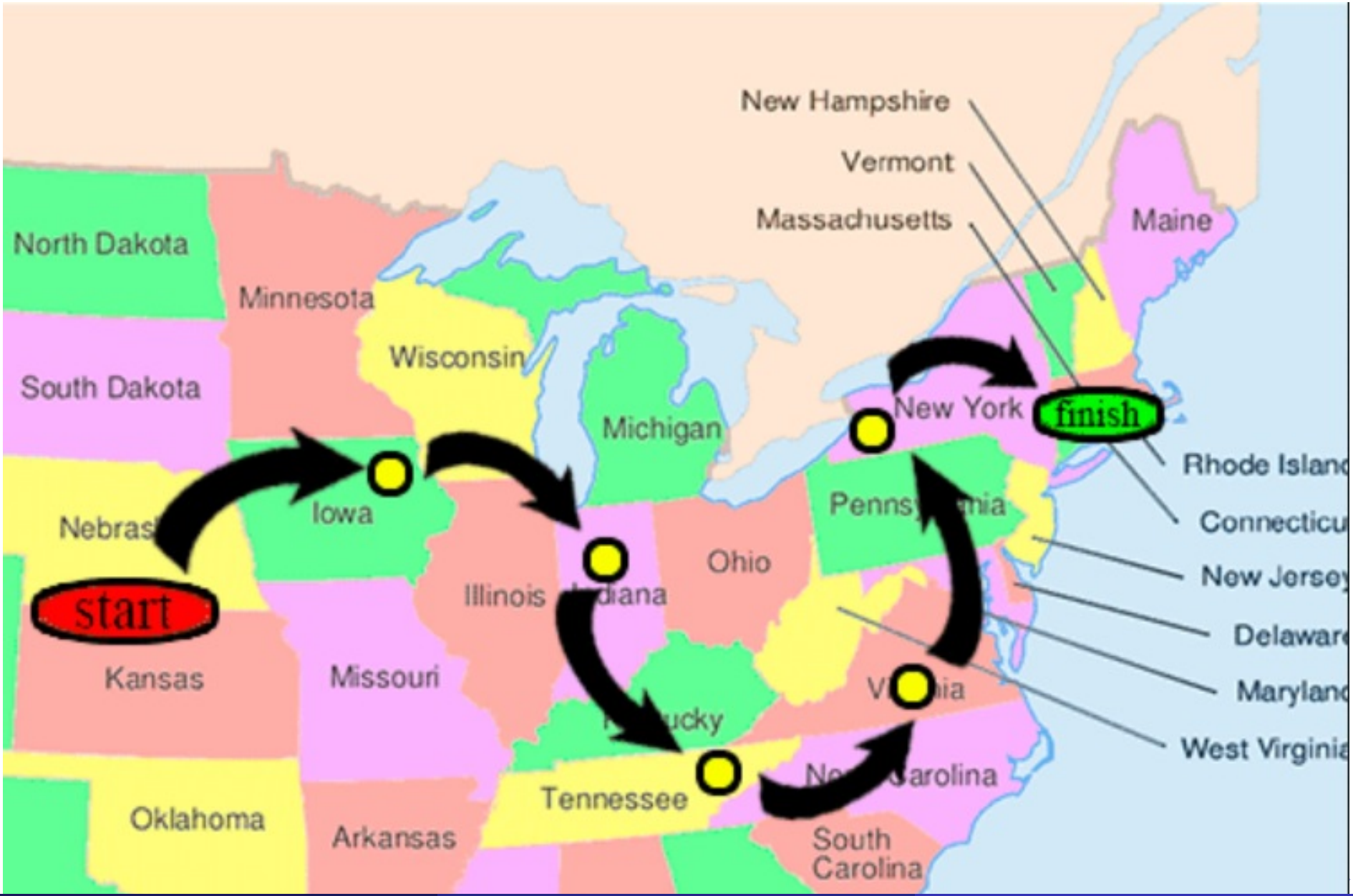
- Python: igraph, graphtool, SNAP и **networkx**
- R: igraph, sna, statnet, ergm, Rsiena. В R доступны сетевые пакеты для статистических сетевых моеделй - exponential random graph models, stochastic actor-oriented models
- Gephi - шикарная визуализация
- ORA - неплохой инструмент с возможностью статистического тестирования гипотез, но естьограничение по объему сети
- Pajek - старый, но работающий инструменты

Эксперимент 'Малый мир'

В 1967 году социолог Стэнли Милгрэм провел эксперимент с целью оценить число связей между двумя случайными людьми.

- Случайным людям из городов Омаха (Небраска) и Уичито (Канзас) были отправлены письма с описанием эксперимента и просьбой отправить письмо *целевому контакту* - человеку в Бостоне (Массачусетс);
- Если участники эксперимента лично знали целевой контакт, они могли отправлять ему письмо напрямую;
- Если участники эксперимента не знали целевой контакт, они должны были выбрать среди своих знакомых того, кто с наибольшей вероятностью был с ним знаком.

Эксперимент 'Малый мир'



Результаты эксперимента 'Малый мир'

- Из 296 стартовых писем финальной цели достигло 64 (29%);
- Цепочка от отправителя до получателя в среднем составила 5.5 человек. В дальнейшем это наблюдение переросло в заключение 'все в мире связаны между собой через шесть рукопожатий';
- Главным фактором для выбора 'посредников' стала географическая близость к целевому контакту.

В дальнейшем проводилось большое число схожих исследований на социальных онлайн-сетях (электронная почта, Facebook, MSN), в которых также было показано, что дистанция между двумя случайными вершинами социальной сети невелика и варьируется между 5 и 6.

Ключевые сетевые показатели

Для описания социальных сетей предложены глобальные и локальные метрики.
Глобальные описывают всю цель в целом. Такие метрики стоит сопоставлять с метриками для аналогичных социальных сетей.
Локальные описывают положение вершин (и ребер) в данной сети. Имеет смысл сопоставлять значения таких метрик для других вершин этой же сети.

- Глобальные метрики: плотность, взаимность, транзитивность
- Локальные метрики: степени центральности

Показатели на уровне графа: плотность

Плотность сети - это отношение числа ребер социальной сети к максимально возможному числу ребер, которые в сети потенциально могли бы быть.
Обычно плотность невелика и достигает нескольких процентов для небольших сетей, и нескольких долей процентов для больших сетей.
$d = \frac{E}{N(N-1)}$,
где d - плотность графа, E - число ребер в сети, N - число вершин в сети.
Укажите, для направленного или ненаправленного графа рассчитана плотность? Каким образом уравнение нужно подправить, чтобы оно выглядело правильно для каждого из типов графов?

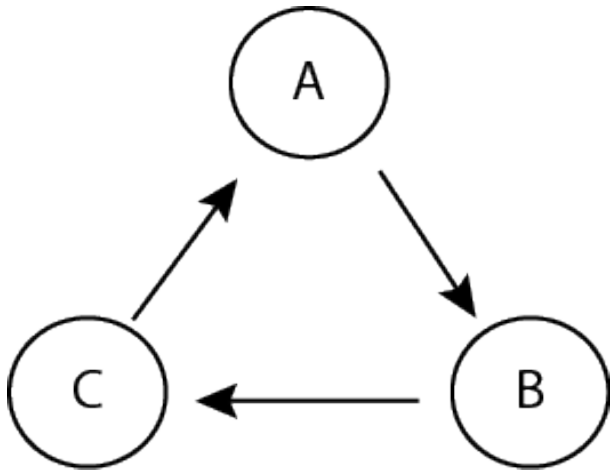
Показатели на уровне графа: взаимность

Показатель взаимности может быть рассчитан только для направленного графа.
Взаимность - это доля взаимных связей ко всем связям.

В большинстве эмпирических сетей показатель взаимности высок, он достигает обычно 0.5.

Показатели на уровне графа: транзитивность

Транзитивность социальной сети показывает склонность к формированию сообществ в сети.
Транзитивность рассчитывается как отноешние закрытых триад ко всем возможным триадам.



Algorithmic confounding

- Почему в социальных сетях очень много сообществ? Почему много друзей друзей? (Ugander, 2010)
- Очень просто - социальная сеть рекомендует друзей на основании алгоритма "друг моего друга мой друг и это является механизмом смещения числа закрытых триад, обусловленного строением социальной сети и рекомендательного алгоритма
- Какой мы из этого можем сделать вывод?

Показатели на уровне вершины: степень центральности

Самая простая и часто используемая метрика, характеризующая локальное положение вершины в социальной сети.
Степень центральности (degree centrality) - число вершин, с которыми соединена данная вершина.
$C_d(i) = deg(i)$
Например, чило друзей Вконтакте - это степень центральности для сети дружбы Вконтакте.
Преимущества - легко считать, легко интерпретировать.
Недостатки - характеризует исключительно локальную позицию вершины и не рассматривает ее в контексте всей сети.

Показатели на уровне вершины: степень близости

Степень близости (closeness centrality) показывает насколько удалена данная вершина от всех остальных вершин сети. Часто эта метрика используется для изучения распространения информации по сетям. Вершины, находящиеся ближе к центру склонны быстрее заражаться, если инфекция происходит из ядра сети.

$C_c(i)=\frac{1}{\sum_j d(i,j)}$

$C_c(i)$ - степень близости вершины i, $d(i,j)$ - расстояние между вершинами i и j.

Чем выше степень центральности, тем меньше расстояние от данной вершины до всех остальных. Соответственно, тем ближе вершина ко всем остальным и тем проще прохождение информации/заражения/ресурсов.

Показатели на уровне вершины: степень посредничества

Степень посредничества (betweenness centrality) показывает уровень "контроля"вершины за распространением информации в сети. В ее основе лежит предположение о том, что

$C_b(i)=\sum_{i\neq j\neq k} \frac{d_{j,k}(i)}{d_{j,k}}$

где $C_b(i)$ - степень посредничества вершины i, $d_{j,k}(i)$ - расстояние между вершинами j и k, если на нем лежит i, $d_{j,k}$ - расстояние между j и k. Чем выше степень посредничества, тем больше сила контроля данной вершины за распространением информации и ресурсов в социальной сети.

Вершины с высокими степенями посредничества называют "брокерами".

Показатели на уровне вершины: эйгенвектор и пейджранк

Эйгенвектор (собственный вектор) - показывает "важность вершины". "Важность"оценивается исходя из социального окружения. "Важная вершина связана с другими важными вершинами"(рекурсивное определение).
Эйгенвектор может быть рассчитан только для ненаправленной социальной сети.
Для направленной сети аналогом эйгенвектора (с небольшими замечаниями) является пейджранк (PageRank). Пейджранк был предложен в 1999 году Пейджем и Брином для ранжирования интернет-страниц.

Посчитаем все центральности на игрушечном графе вручную