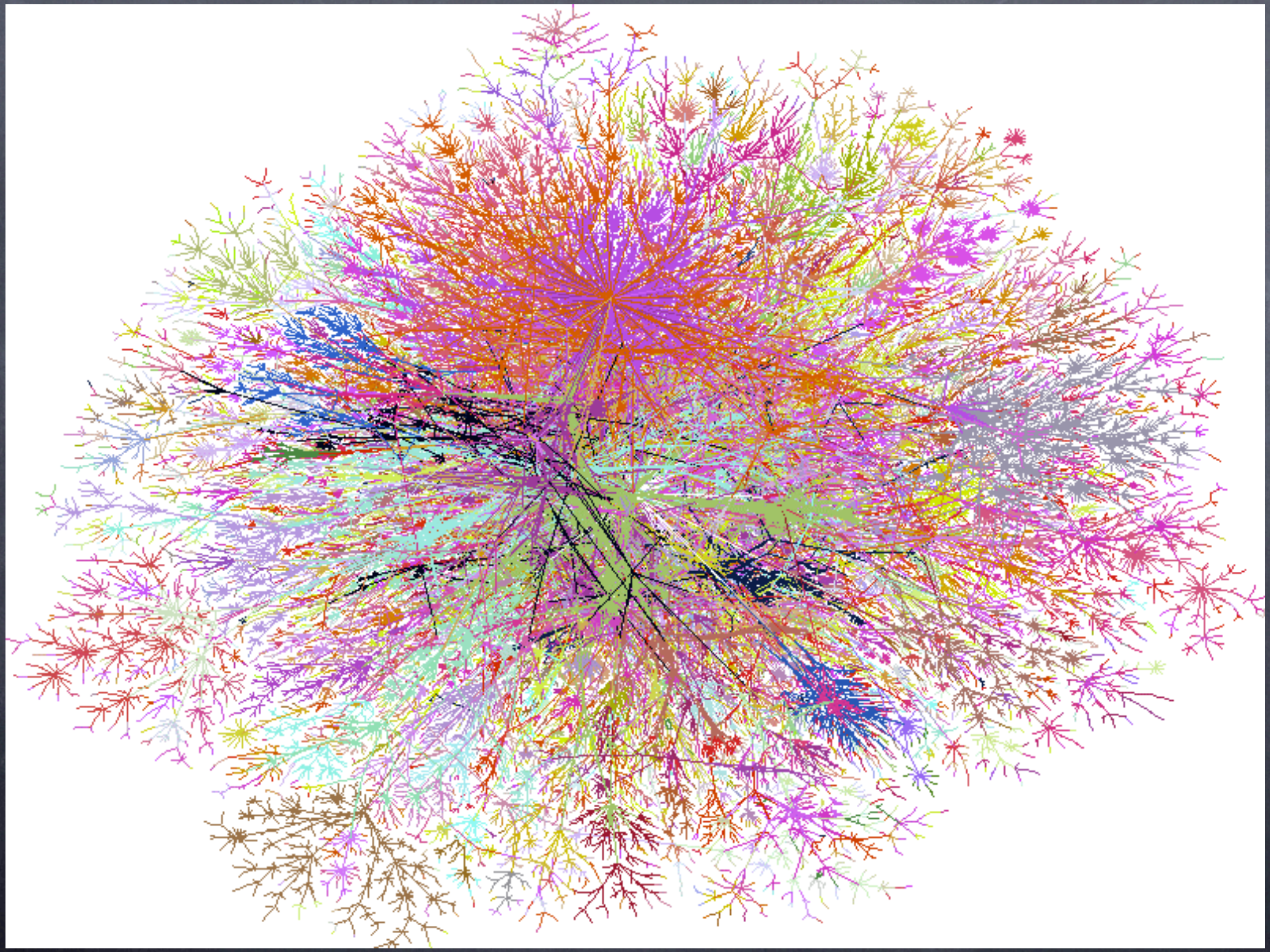
Анализ Безмашстабных Сетей

Павел Велихов, ИСП РАН
Московская Секция ACM SIGMOD
29.03.2007



Что описывает?

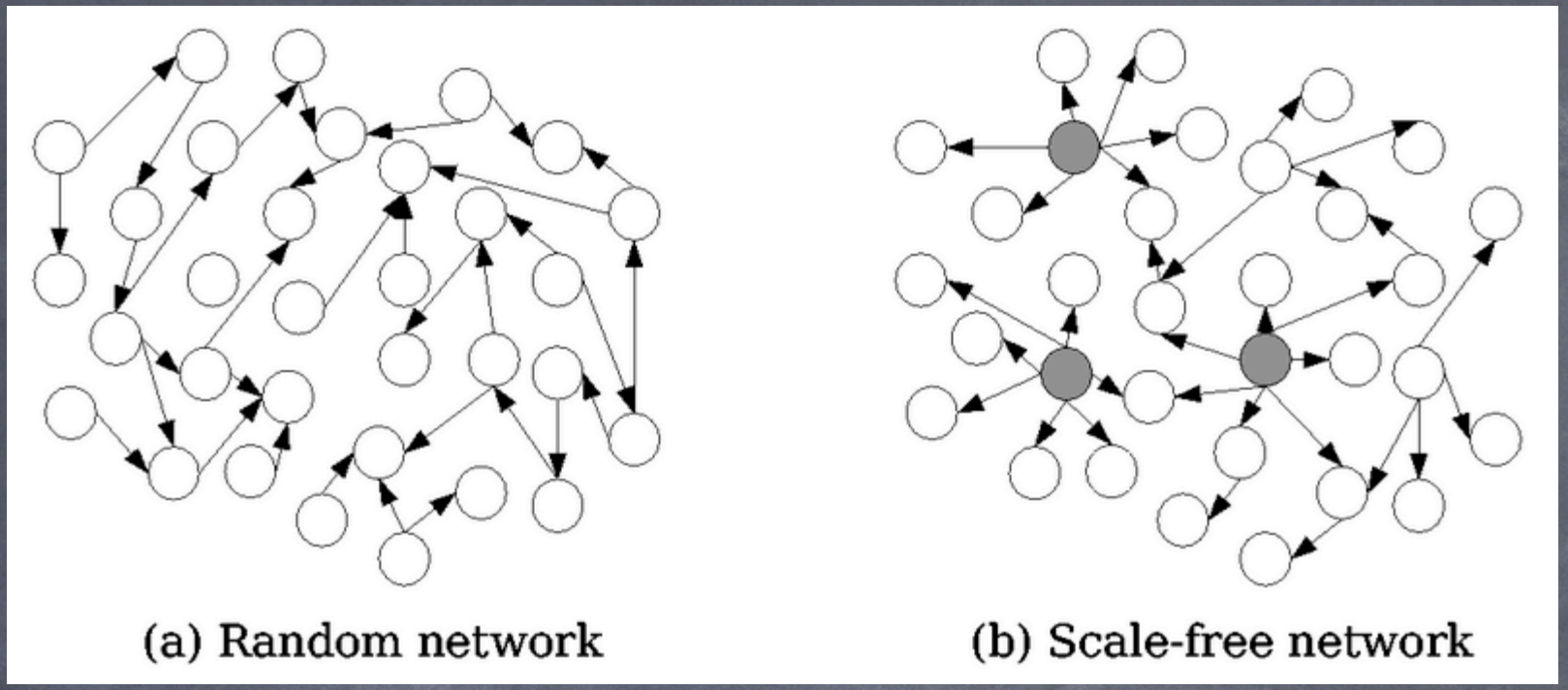
- Техника: Сети электропередачи, VLSI, Интернет, Веб, ...
- Социум: контакты, связи, организации, язык, дороги, авиамаршруты, ...
- Биология: нейроны; пищевые, экологические, метаболические сети, ...
- Физика: молекулы, галактики

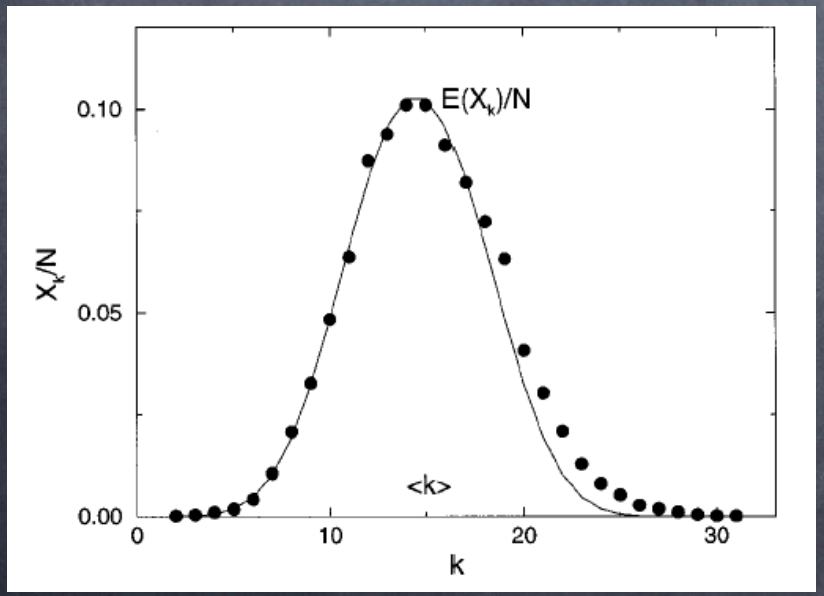
План доклада

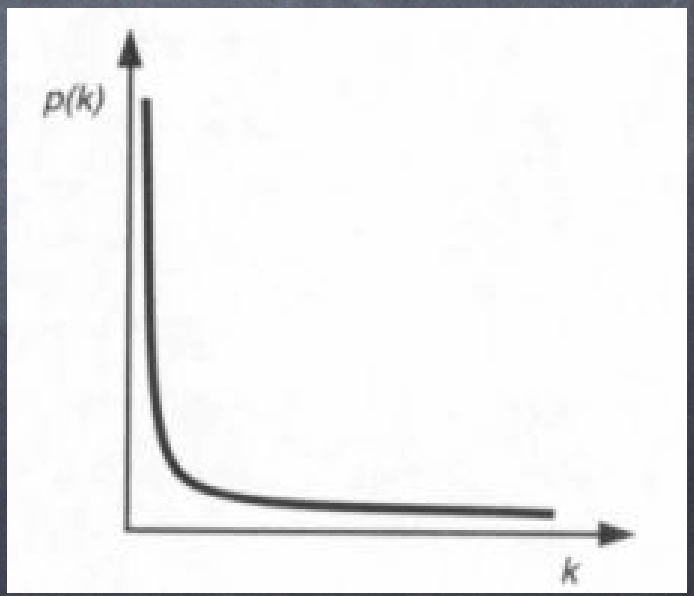
- Основные тезисы: структура, эволюция, динамика
- Успехи теории: вирусология, социология, траспортные сети
- Применения в построении систем: P2P, SpamRank
- Методы анализа сетей и новые требования к СУБД

Универсальная теория

- Структура
  - Хабы, малый мир, устойчивость, кластеризация
- Эволюция
  - Предпочтительное соединение
- Динамика
  - Быстрая реакция, Нулевой порог в эпидемологии, ...


(a) Random network
(b) Scale-free network


E(X_k)/N
X_k/N
0.10
0.05
0.00
0
10
20
30
<k>
k

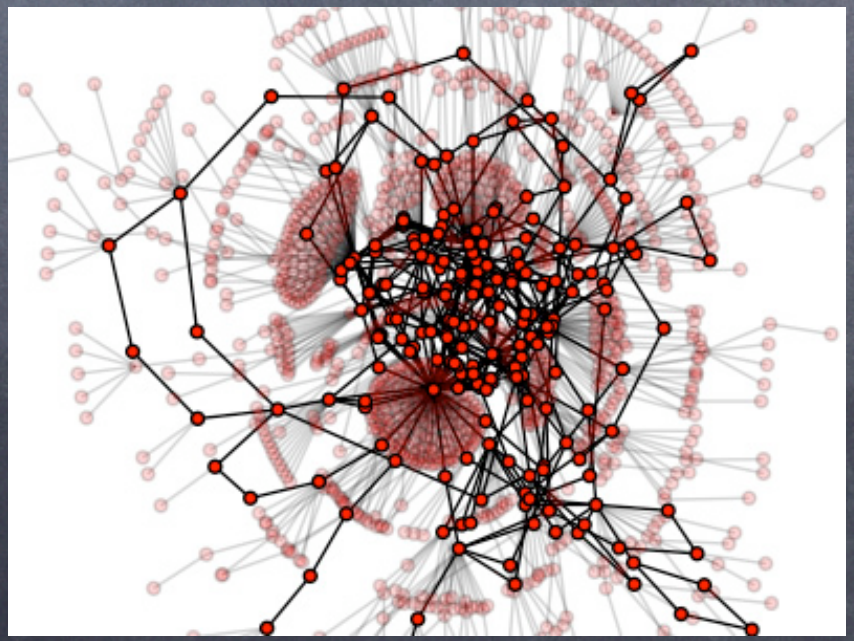

p(k)
k

Структура сетей

- Распределение: $P(k) \sim k^{-\gamma}$
  - Степенной закон: $P(k>a \mid k>b) = (a/b)^{-\gamma}$
- Дополнительные метрики:
  - Транзитивность: $P(e(a,c) \mid e(a,b), e(b,c))$
  - ...
  - S-метрика: $\Sigma\, d_i d_j$ если $e(i,j)$ [4]

Структурные свойства

- Хабы, которые связывают сеть, большое ядро, малый диаметр - ln(ln(N)) [3]



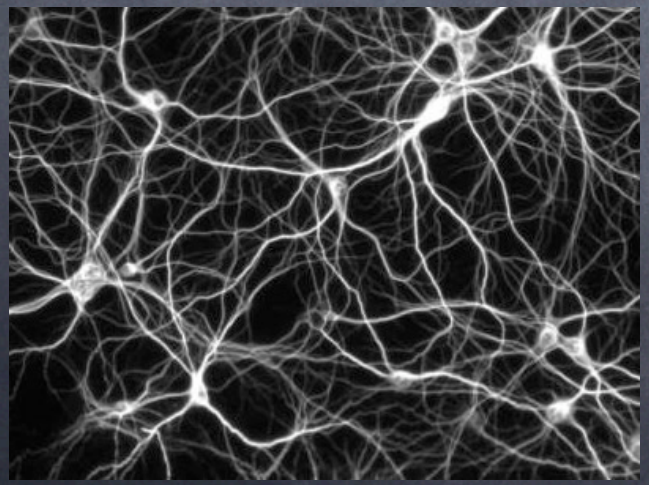
Структурные свойства

- Устойчивость к случайным разрушениям, чувствительность к аттакам

Случайное разрушение (удаляем случ. узлы)	
Случайный граф – диаметр растет линейно	Безмасштабные сети – диаметр не растет
Аттака (удаляем максимальные узлы)	
Случайный граф – диаметр растет линейно	Безмасштабные сети – диаметр удваивается после каждых 5% узлов

Структурные свойства

- Само-похожесть, свойства не зависят от размера сети
  - Самопожесть из-за распределения
  - Динамические свойства зависят от γ



Механизмы образования

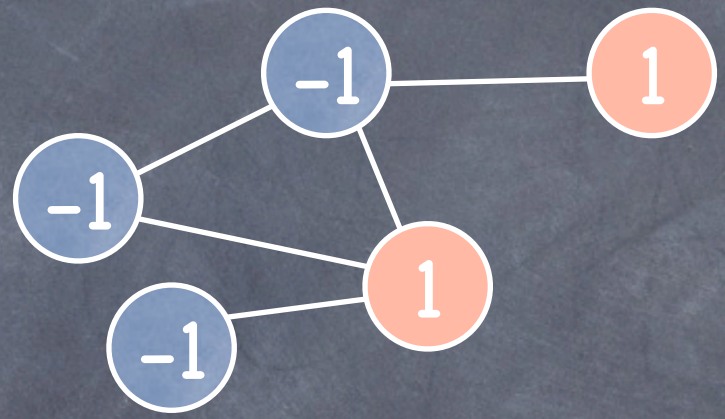
- Предпочтительное присоединение
  - интернет, актеры, цитируемость публикаций, авиа маршруты
- Дополнительные факторы: асортативность, ...

Пока все про топологию

- Пока смотрели только на сети с равноправными ребрами
- На практике важнее взвешенные сети
- Очень часто высокая степень узла коррелирует с высоким весом

Динамические свойства

- Дискретные процессы[5] - модель Изинга



- $f_{t+1}(x) = \text{sign}(\Sigma\ f_t(\Gamma(x)))$
- существует fixpoint, +1 или -1
- Метод среднего поля $\partial F_k(t) = ...F_k(t)...$

Динамические свойства

- Если $\Upsilon > 5/2$, $t(\text{fixpoint}) = O(\log(N))$
- Если $\Upsilon < 5/2$, $t(\text{fixpoint}) = \text{const}$

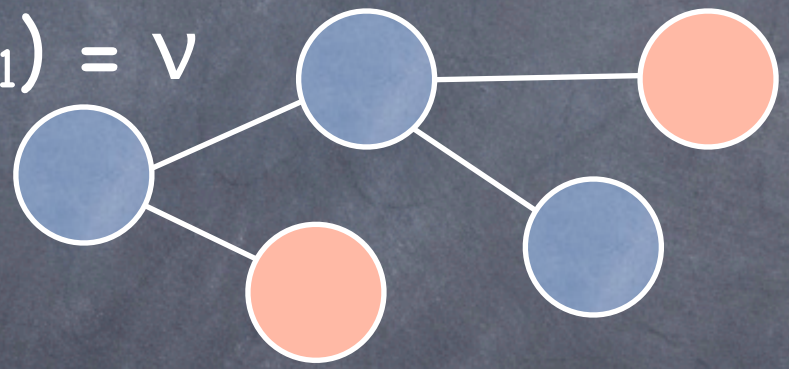
План доклада

- Основные тезисы: структура, эволюция, динамика
- Успехи теории: вирусология, социология, коммуникации
- Применения в построении систем: P2P, SpamRank
- Методы анализа сетей и новые требования к СУБД

Вирусология[6]

- SIS модель: здоров – болен – здоров

$P(I_{t+1}) = \nu$



$P(S_{t+1}) = \delta$

- $\lambda = \nu/\delta$
- метод среднего поля
- $\lambda < \lambda_0$ – вирус затухает

Вирусология

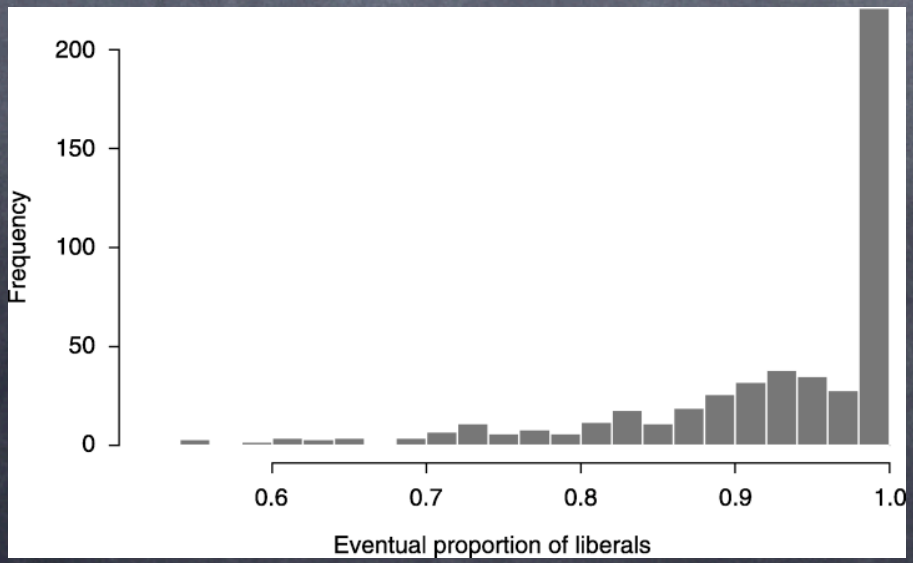
- Для безмасштабных сетей $\gamma < 3$, $\lambda_0 = 0$
- Лечить случайные узлы ни к чему не приводит
- Если лечить хабы, то можно быстро поднять λ_0
- СПИД, компьютерные вирусы

Социология[7]

- Исследовали социальные сети, только не безмасштабные
- Hayek: как социализм/коммунизм распространился в западных странах в первой половине 20-ого века?
- "In all democratic countries, a strong belief prevails that the influence of the intellectuals on politics is negligible."

Социология

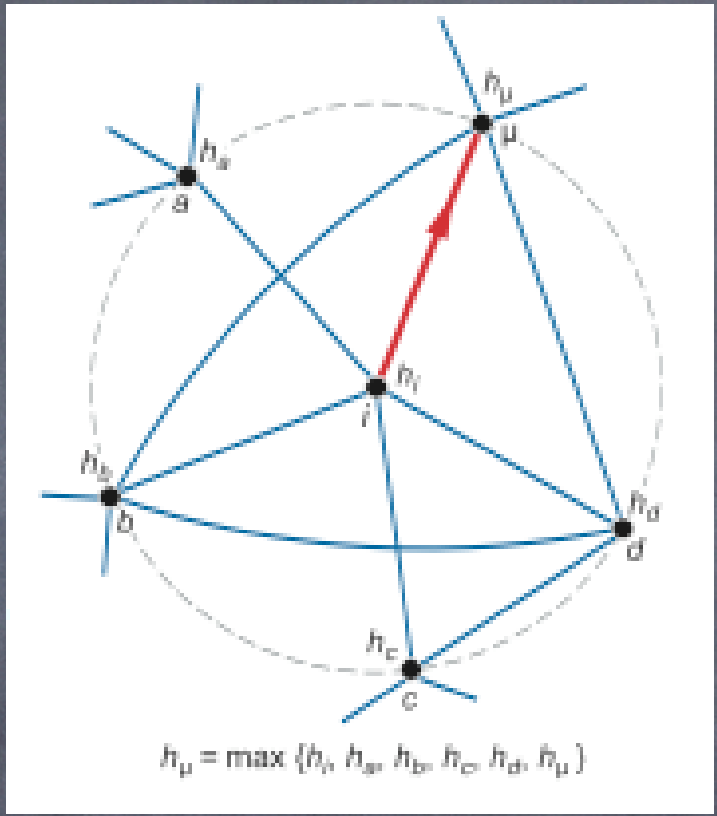
- Дискретная модель влияния: 4 соседа, взвешенные по степени
- 80% консервативные, 2% максимальных узлов социалисты



Траспортные сети

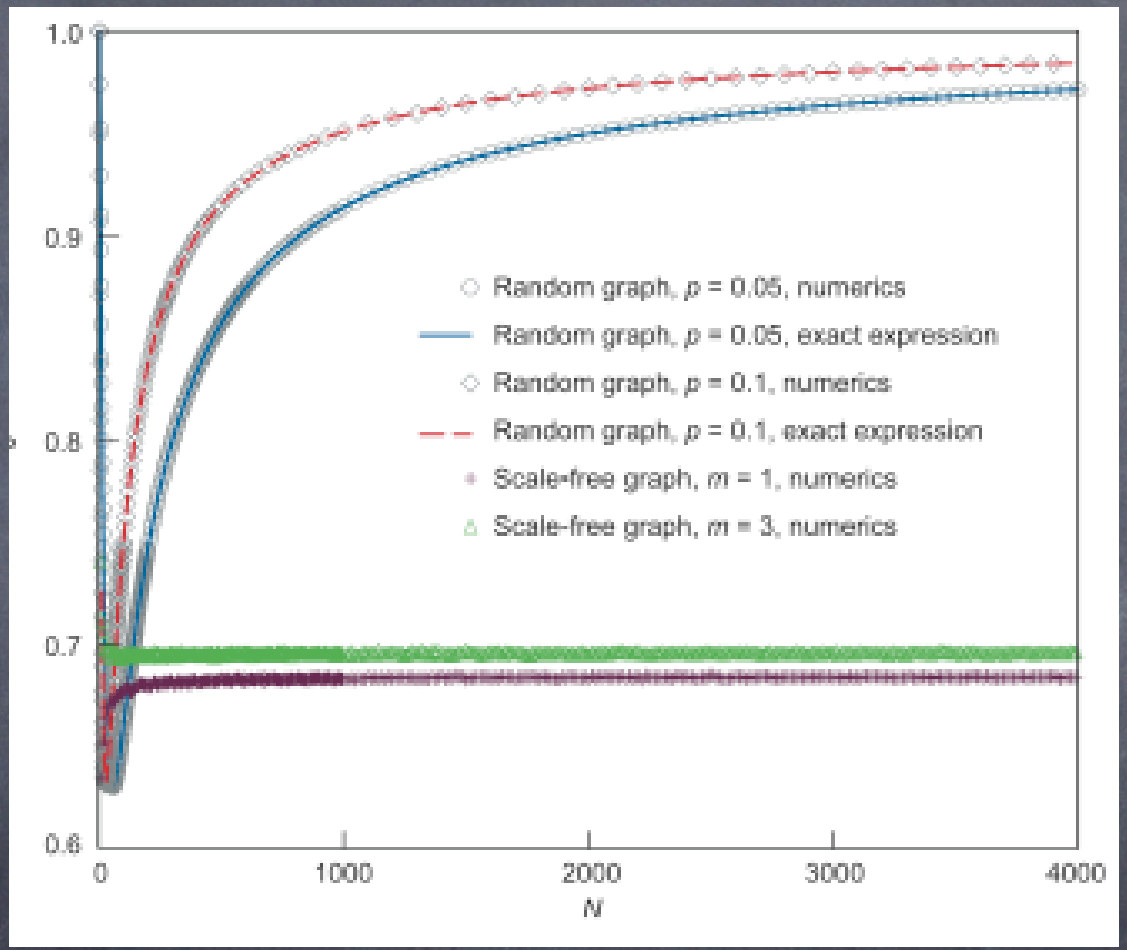
- Интернет, авиа-маршруты, дороги, электропередача
- Как топология влияет на заторы?
- Градиентный граф

Транспортные сети[8]



- Градиентный граф все время меняется
- Но можно посчитать $J = 1 - N_{in}/N_{out}$
- Безмасштабная сеть – J не зависит от размера сети

Транспортные сети

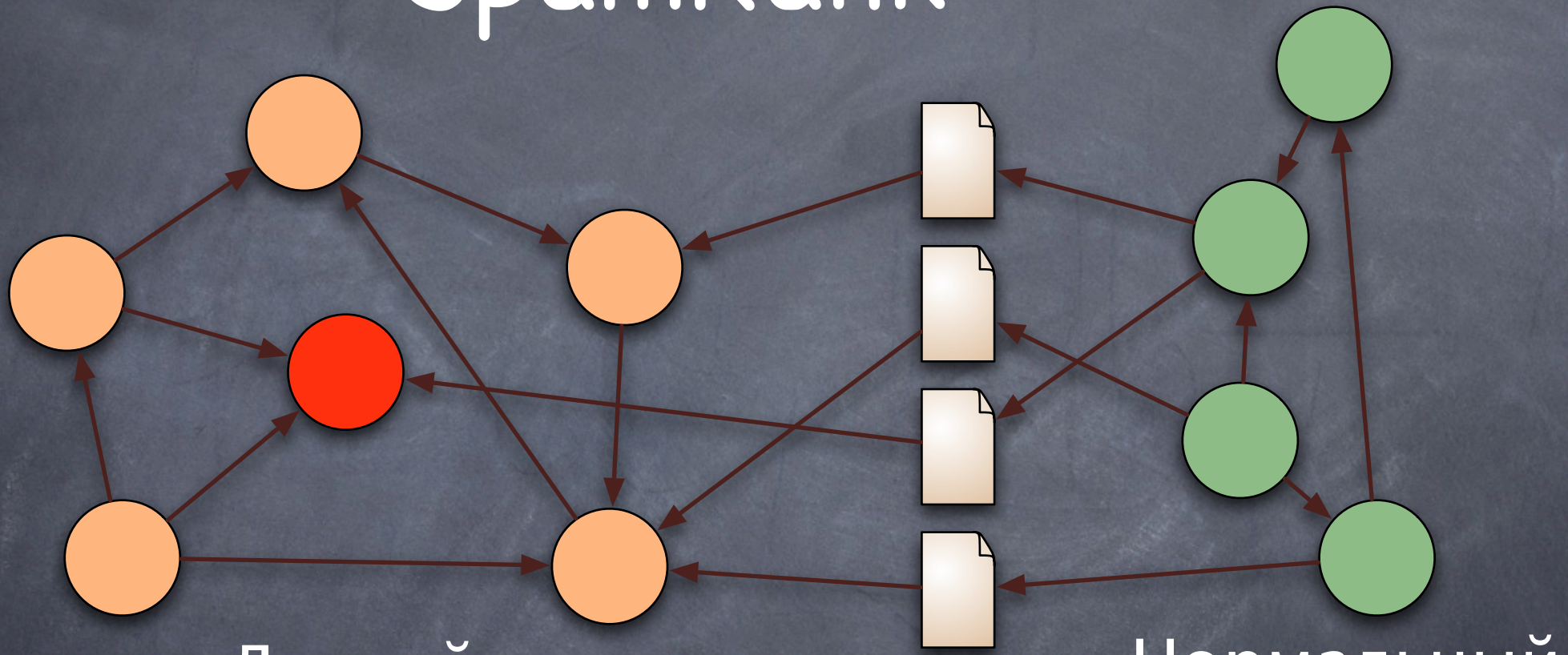


- В пределе, в безмасштабном графе заторов на порядок меньше

План доклада

- Основные тезисы: структура, эволюция, динамика
- Успехи теории: вирусология, социология, траспортные сети
- Применения в построении систем: P2P, SpamRank
- Методы анализа сетей и новые требования к СУБД

SpamRank[9]



SpamRank

- Ищем подозрительное распределение рангов в округе узла с высоким PageRank
- Support - собираем методом Монте-Карло, random walk от узла с затуханием
- Сравниваем распределение PageRank с нормальным внутри Support

P2P сети

- Случайным образом выросли в безмасштабные сети
- Обычный поиск приводит к flood
- Направленный поиск в сторону узлов с максимальной степенью
- O(logN) [10]

План доклада

- Основные тезисы: структура, эволюция, динамика
- Успехи теории: вирусология, социология, траспортные сети
- Применения в построении систем: P2P, SpamRank
- Методы анализа сетей и новые требования к СУБД

Что нужно от СУБД?

- Хранение, поиск
- Простые метрики - распределение, диаметр, корреляция
- Запросы хитрее: сообщества, похожесть по распределению, потоки
- Совсем страшно: темпоральный анализ, предсказание

Хранение

- Кластеризация: запросы к сетям будут бегать по соседям. Большинство узлов можно хранить с соседями 1-2-3 порядка
- Что делать с хабами? С костяком?
- Зависит от кластеризации или s-метрики.

Поиск

- k-NN: A* или другой поиск может попасть в костяк сети
- Индексы: метрика расстояния $d(a,b) = \Sigma w_i \in path(a,b)$ не удволетровяет неравенству треугольника
- Статистические алгоритмы: P, ϵ

Простные метрики

- Диаметр
- Корреляция, s-метрика
- Можно ли считать по случайной выборке?

Хитрые запросы

- Сообщества: Σ in-link >> Σ out-link, ищутся используя иерархическую кластеризацию [11]
- Потоки: максимальная пропускная способность, т.п. Стандартные алгоритмы не учитывают топологию

Хитрые запросы

- Похожесть по свойствам окружения[12]
- $sim(a,a) = 1$
- $sim(a,b) = \alpha * \Sigma\Sigma\, sim(I_i(a),I_j(b)) / |I(a)||I(b)|$, α – фактор затухания
- как адаптировать к безмасштабным графам?

Темпоральный анализ

- “Покажи мне сайты, которые развивались как Google”
- “Найди пузыри в экономике”

Предсказание

- Кто с кем напишет следующую статью по базам данных?
- Где проложат новые дороги?
- Кто присоеденится к группе риска?

Link prediction

- Модель предсказания: контраст с supervised learning

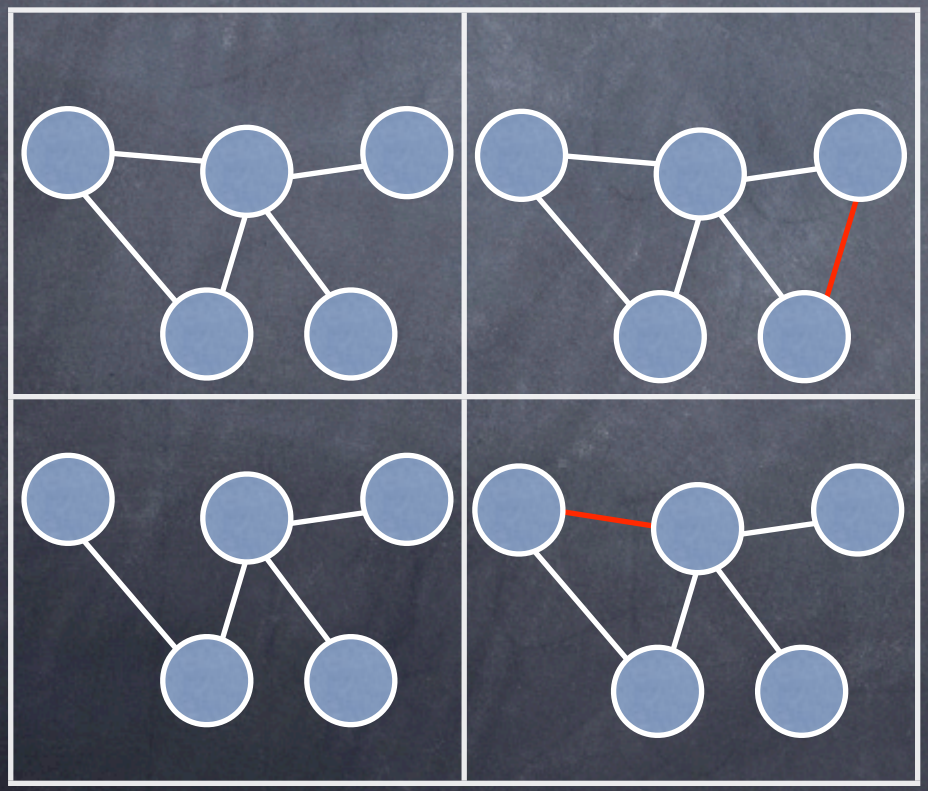
A_1	A_2	A_3	C
a_{11}	...	...	1
a_{12}	...	...	0
a_{12}	...	...	1
a_{13}	...	...	0

$f(A_1 \ldots A_n) \rightarrow C$

Link Prediction

- Модель предсказания связей

E_t, V_t E_{t+1}, V_t



$f(E,V) \rightarrow E$

Оценка:

$top_n[\ f(E_t,V_t)\ ] \cap E_{t+1},$

$n = |E_{t+1}|$

Link Prediction[13]

- Результаты:
  - со-авторство у физиков
  - 8 - 10% правильных предсказаний в сетях со-авторства, лучший метод - общие соседи: score (x,y) = # общих соседей
  - 0.1 - 0.4% базовый предсказатель

Что нужно от СУБД?

- Хранение, поиск
- Простые метрики - распределение, диаметр, корреляция
- Запросы хитрее: сообщества, похожесть по распределению, потоки
- Совсем страшно: темпоральный анализ, предсказание

Библиография

- Не покрыли:
  - Huang, Z. "Link Prediction Based on Graph Topology: The Predictive Value of the Generalized Clustering Coefficient"
  - Hein, Oliver; Schwind, Michael; Spiwoks, Markus. "A Microscopic Stock Market Model with Heterogeneous Interacting Agents in a Scale-Free Communication Network"
- Много других статей, не упомянутых в абстракте
- TBD