Моделирование социальных сетей

Даниил Мусатов

Московский физико-технический институт

21 июля 2012

Что такое социальная сеть

Социальная сеть состоит из:

- Агентов;
- Связей между агентами.

Как правило, связи двусторонние, но могут быть и односторонними.

Что такое социальная сеть

Социальная сеть состоит из:

- Агентов;
- Связей между агентами.

Как правило, связи двусторонние, но могут быть и односторонними.
Традиционные обозначения: V — множество агентов, E — множество связей.

Примеры социальных сетей

- Контакты между отдельными людьми:
  - Личные
  - Онлайновые
  - Деловые
  - Профессиональные
  - Криминальные
  - ...

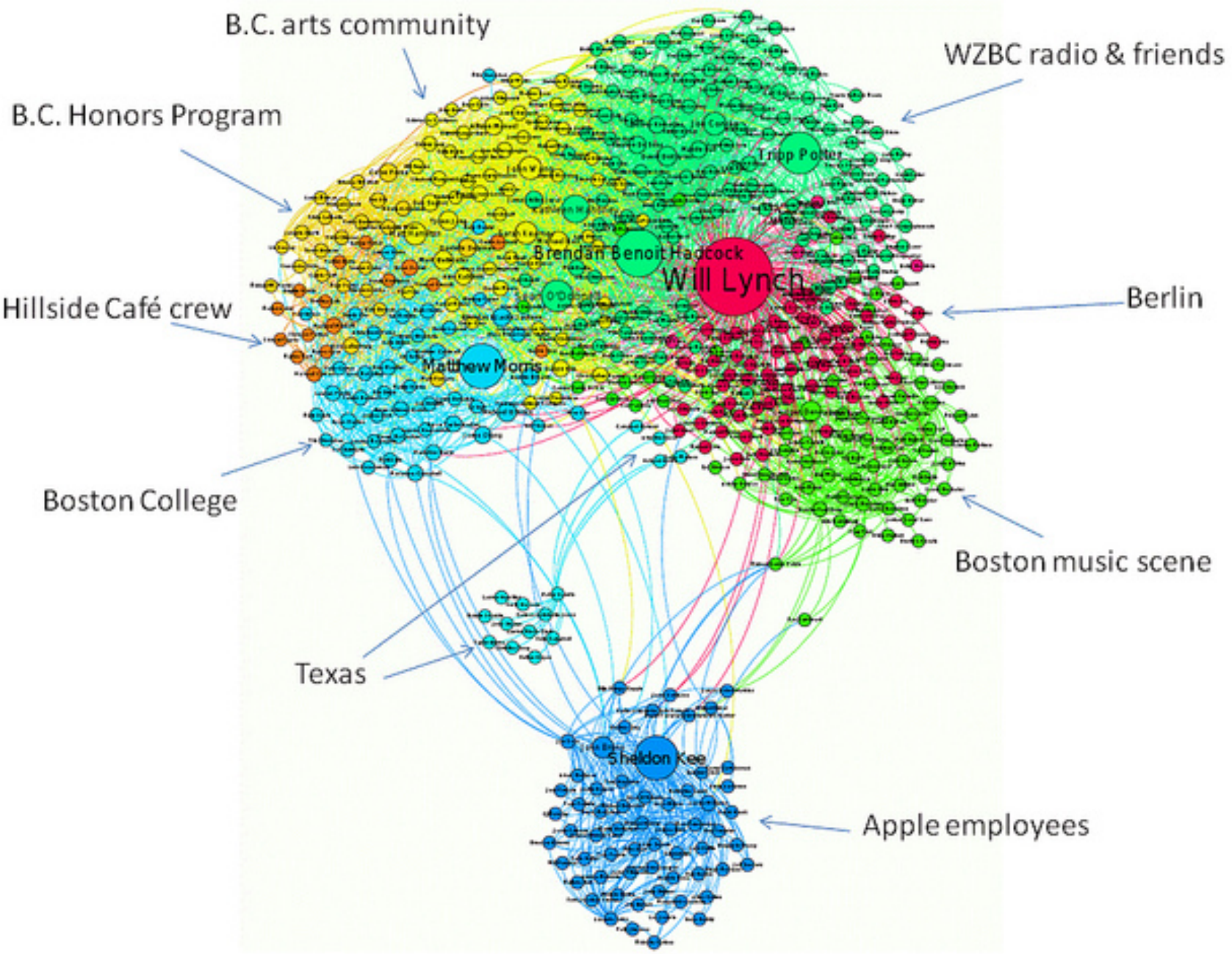
Примеры социальных сетей

- Контакты между отдельными людьми:
  - Личные
  - Онлайновые
  - Деловые
  - Профессиональные
  - Криминальные
  - ...
- Сети совместной деятельности: есть группы людей, занимающихся общим делом, связи проводятся между людьми, участвующими в одном проекте:
  - Соавторство научных статей
  - Съёмки в одних фильмах
  - Членство в онлайн-сообществах
  - ...

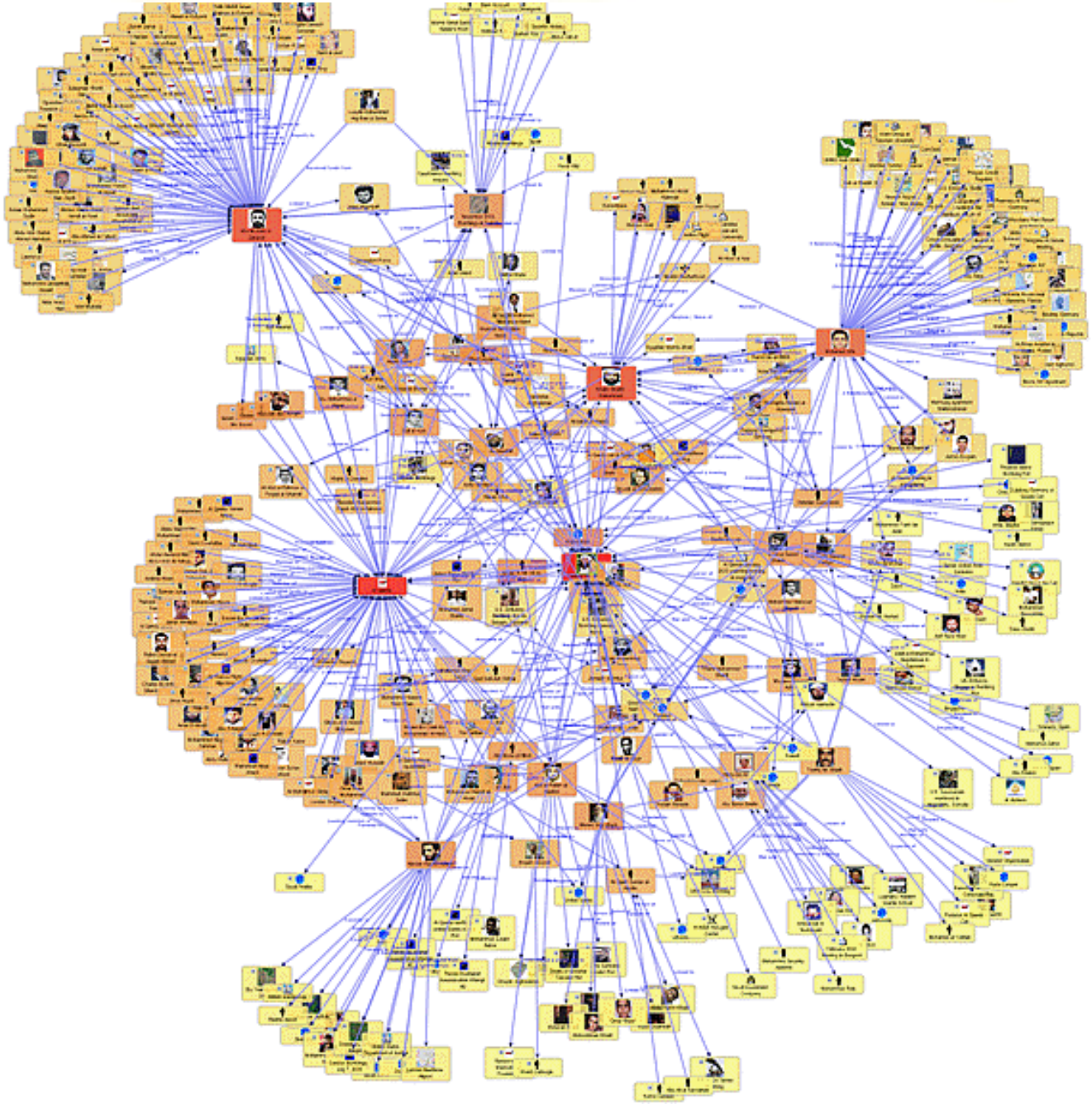
Примеры социальных сетей

- Контакты между отдельными людьми:
  - Личные
  - Онлайновые
  - Деловые
  - Профессиональные
  - Криминальные
  - ...
- Сети совместной деятельности: есть группы людей, занимающихся общим делом, связи проводятся между людьми, участвующими в одном проекте:
  - Соавторство научных статей
  - Съёмки в одних фильмах
  - Членство в онлайн-сообществах
  - ...
- Двудольные сети: мужчины-женщины, авторы-статьи, фирмы-работники и т.п.

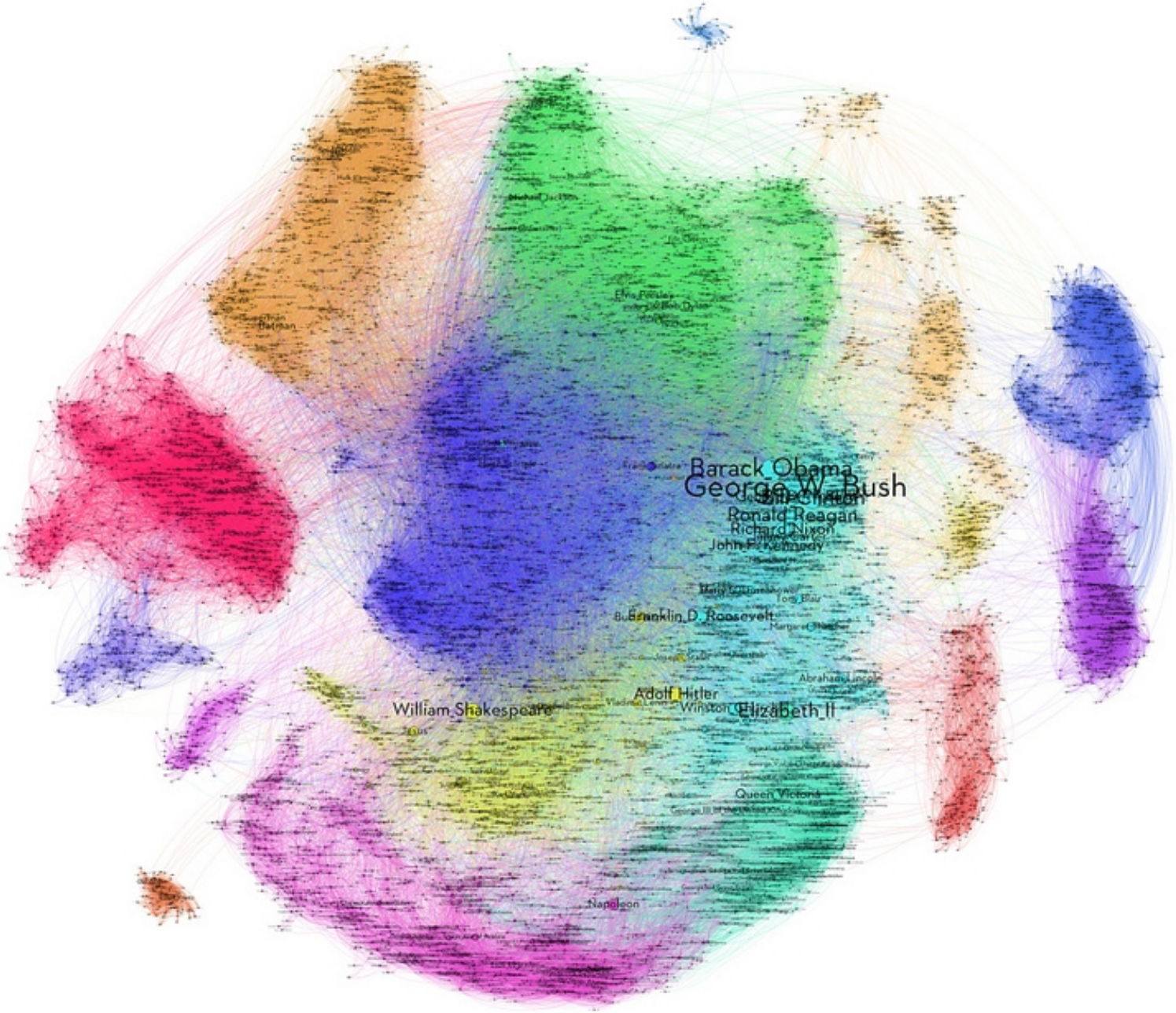
Фрагмент facebook



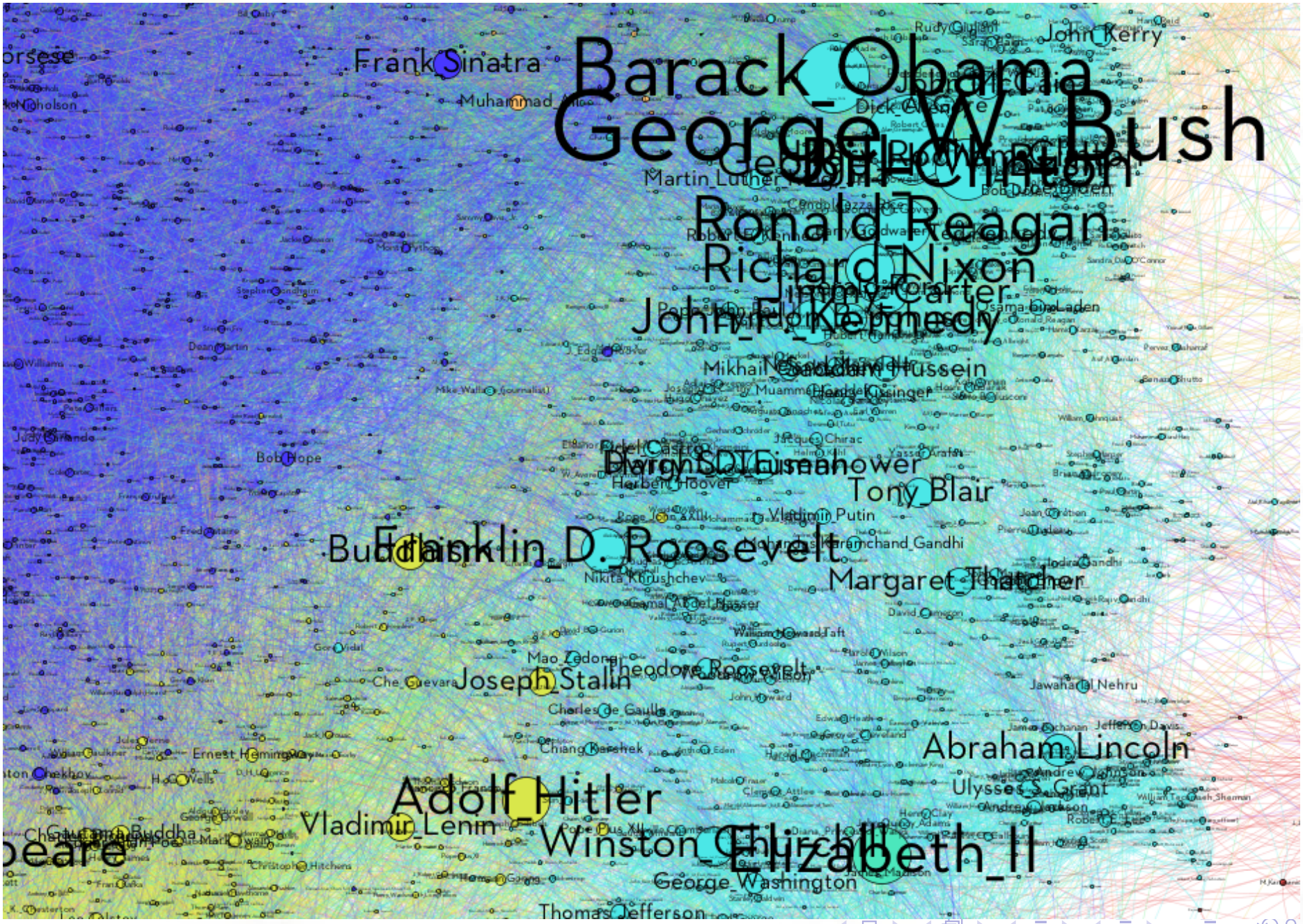
Террористы, осуществившие атаки 11 сентября



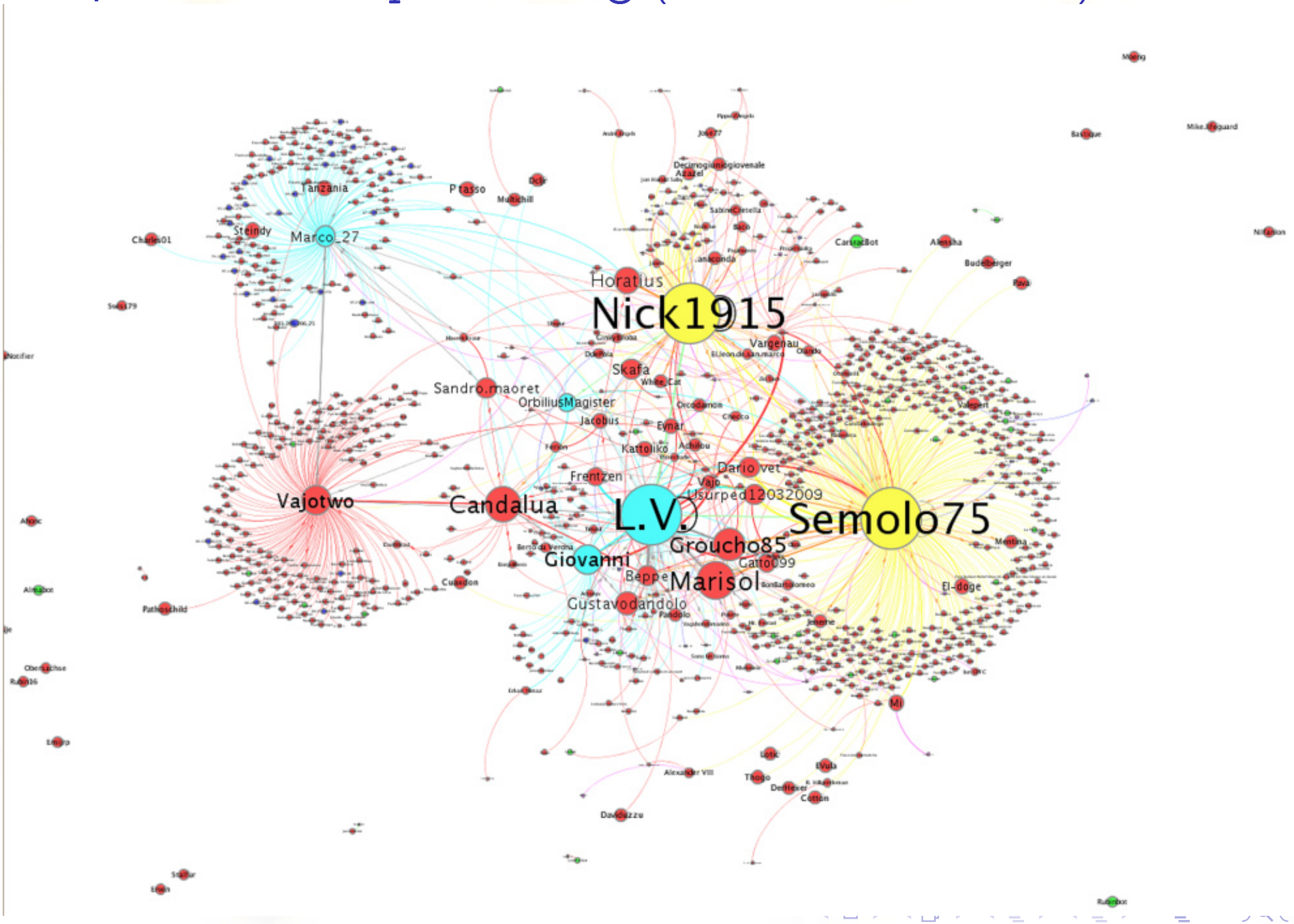
Персонажи статей википедии (hackdiary.com)



Персонажи статей википедии (фрагмент)



Авторы vec.wikipedia.org (Paolo Massa, 2011)



Числовые характеристики узлов и сети в целом

- Расстояние между двумя вершинами
- Диаметр графа
- Степень вершины (количество контактов)
- Распределение степеней вершины
- Меры центральности узла (closeness centrality и betweenness centrality)
- Распределение меры центральности
- Коэффициент кластеризации
- Коэффициент ассортативности

Теория шести рукопожатий

Утверждение

Между любыми двумя людьми на земле можно установить цепочку знакомств средней длины 6.

Расстояния в графе

- *Расстояние между двумя вершинами* — длина кратчайшего пути, соединяющего эти две вершины
- *Диаметр графа* — максимальное расстояние между двумя вершинами
- *Радиус графа* — $\min_v \max_w \mathrm{dist}(v, w)$
- **Переформулировка утверждения**: в графе знакомств жителей Земли среднее расстояние между двумя вершинами равно 6.

Теория шести рукопожатий: история вопроса

1929 — Фридеш Каринти, рассказ «Звенья цепи»

A fascinating game grew out of this discussion. One of us suggested performing the following experiment to prove that the population of the Earth is closer together now than they have ever been before. We should select any person from the 1.5 billion inhabitants of the Earth – anyone, anywhere at all. He bet us that, using no more than five individuals, one of whom is a personal acquaintance, he could contact the selected individual using nothing except the network of personal acquaintances.

Теория шести рукопожатий: история вопроса

1967 — эксперимент Стэнли Милгрома (“small world experiment”)

- Случайно выбранным жителям Омахи (Небраска) и Уичиты (Канзас) предлагалось переслать письмо адресату в Бостоне, используя только личные знакомства
- Из 296 писем 64 достигли цели
- Длины цепочек варьировались от 2 до 10
- Средняя длина цепочки — около 6

Теория шести рукопожатий: история вопроса

1967 — эксперимент Стэнли Милгрома (“small world experiment”)

- Случайно выбранным жителям Омахи (Небраска) и Уичиты (Канзас) предлагалось переслать письмо адресату в Бостоне, используя только личные знакомства
- Из 296 писем 64 достигли цели
- Длины цепочек варьировались от 2 до 10
- Средняя длина цепочки — около 6

1969 — Casper Goffman: *And what is your Erdős number?*

Теория шести рукопожатий: история вопроса

1967 — эксперимент Стэнли Милгрома (“small world experiment”)

- Случайно выбранным жителям Омахи (Небраска) и Уичиты (Канзас) предлагалось переслать письмо адресату в Бостоне, используя только личные знакомства
- Из 296 писем 64 достигли цели
- Длины цепочек варьировались от 2 до 10
- Средняя длина цепочки — около 6

1969 — Casper Goffman: *And what is your Erdős number?*

- Mine is three: Musatov–Shen–Alon–Erdős

Теория шести рукопожатий: история вопроса

1967 — эксперимент Стэнли Милгрома (“small world experiment”)

- Случайно выбранным жителям Омахи (Небраска) и Уичиты (Канзас) предлагалось переслать письмо адресату в Бостоне, используя только личные знакомства
- Из 296 писем 64 достигли цели
- Длины цепочек варьировались от 2 до 10
- Средняя длина цепочки — около 6

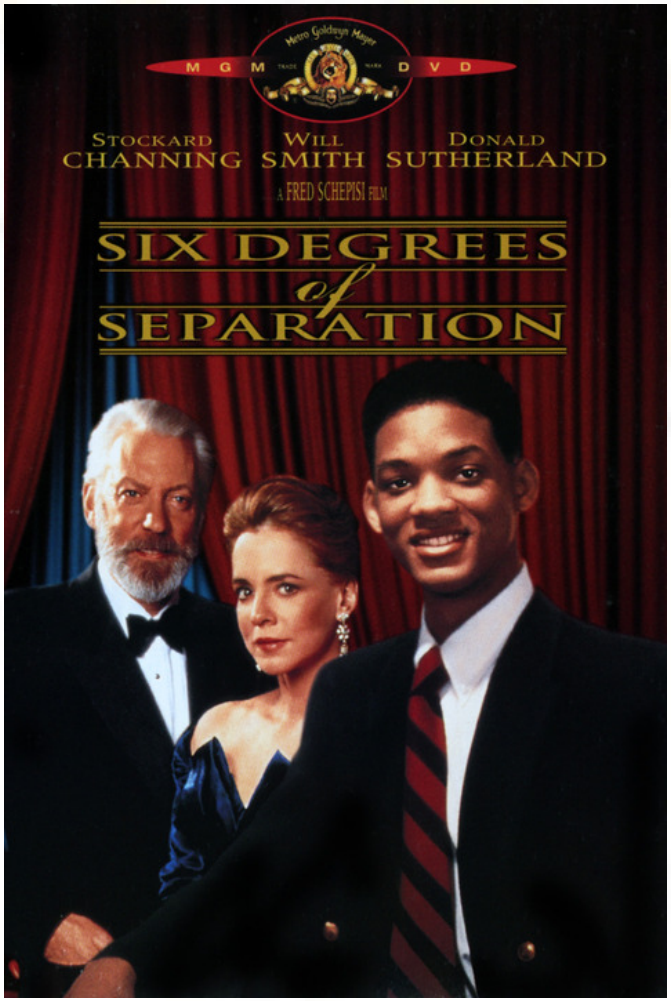
1969 — Casper Goffman: *And what is your Erdős number?*

- Mine is three: Musatov–Shen–Alon–Erdős
- Or Stiglitz number?

Теория шести рукопожатий: история вопроса

1990 — выход пьесы “Six degrees of separation”

1993 — выход одноимённого фильма



Теория шести рукопожатий: проверка на онлайн-сетях

2003 — Повторение эксперимента Милгрома Уоттсом и Строгатцом с использованием электронной почты.

- Среднее расстояние около 6
- Большой процент потерь

2006 — Исследование сообщений в MSN Messenger (LesKovec, Horvitz), среднее расстояние 6,6

2006 — Исследование сетей Orkut, Livejournal, Flickr, Youtube (Mislove et al), среднее расстояние 4.25–5.88, диаметр графа 9–27

2010 — Исследование Twitter (Alex Cheng), среднее расстояние 4.67

2011 — Исследование Facebook (Backstrom et al), среднее расстояние 4.74

Распределение степеней вершин и безмасштабные сети

- Степень вершины — количество её соседей
- Распределение степеней: $P(k) = \#\{v \mid \deg v = k\}$
- Независимость от масштаба (scale-freeness): $P(k) \approx ck^{-\gamma}$
- Эмпирическая проверка: функция распределения степеней в log-log-масштабе должна быть примерно прямой линией.

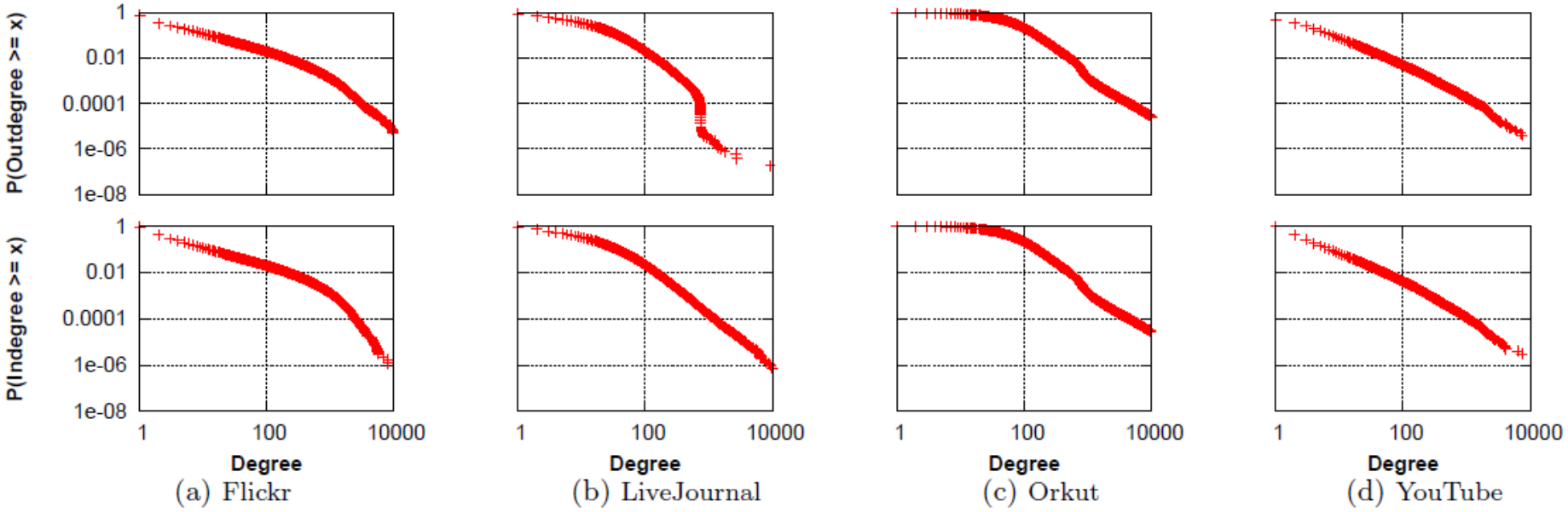
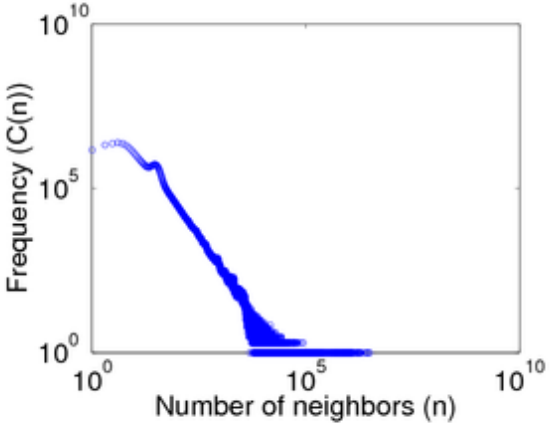


Figure 2: Log-log plot of outdegree (top) and indegree (bottom) complementary cumulative distribution functions (CCDF). All social networks show properties consistent with power-law networks.

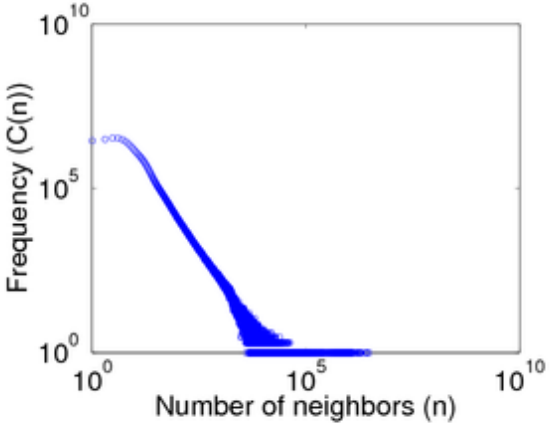
Источник: Mislove et al, 2006

Распределение степеней вершин: примеры

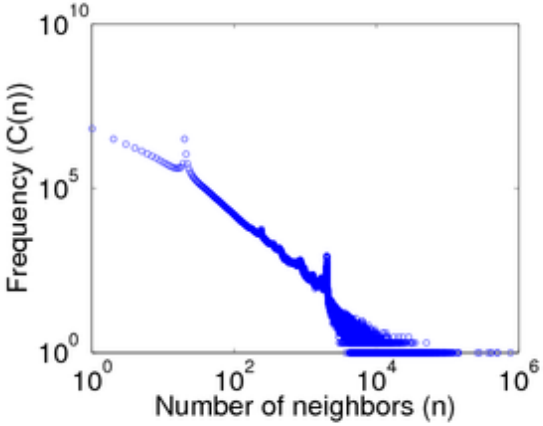
Twitter:



Degree distribution

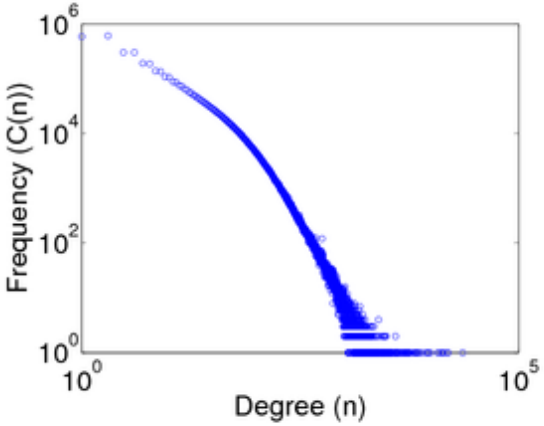


Outdegree distribution

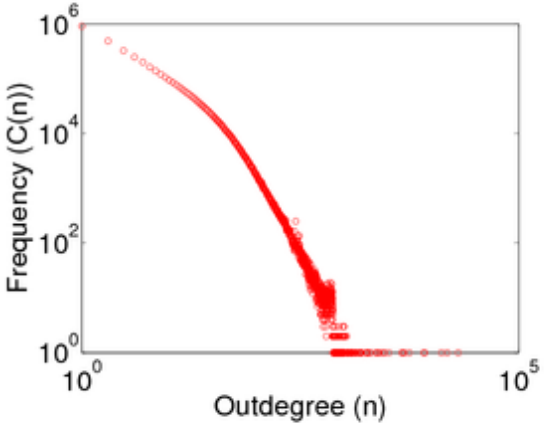


Indegree distribution

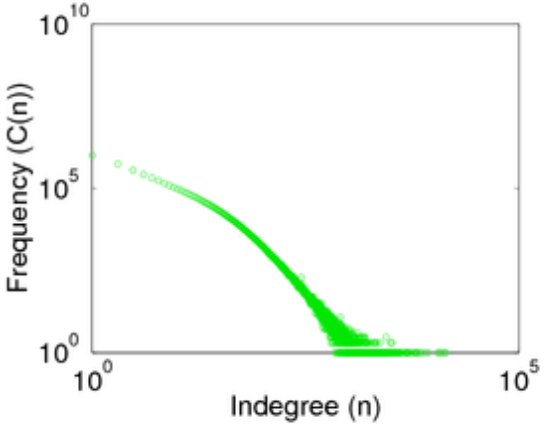
Livejournal:



Degree distribution



Outdegree distribution

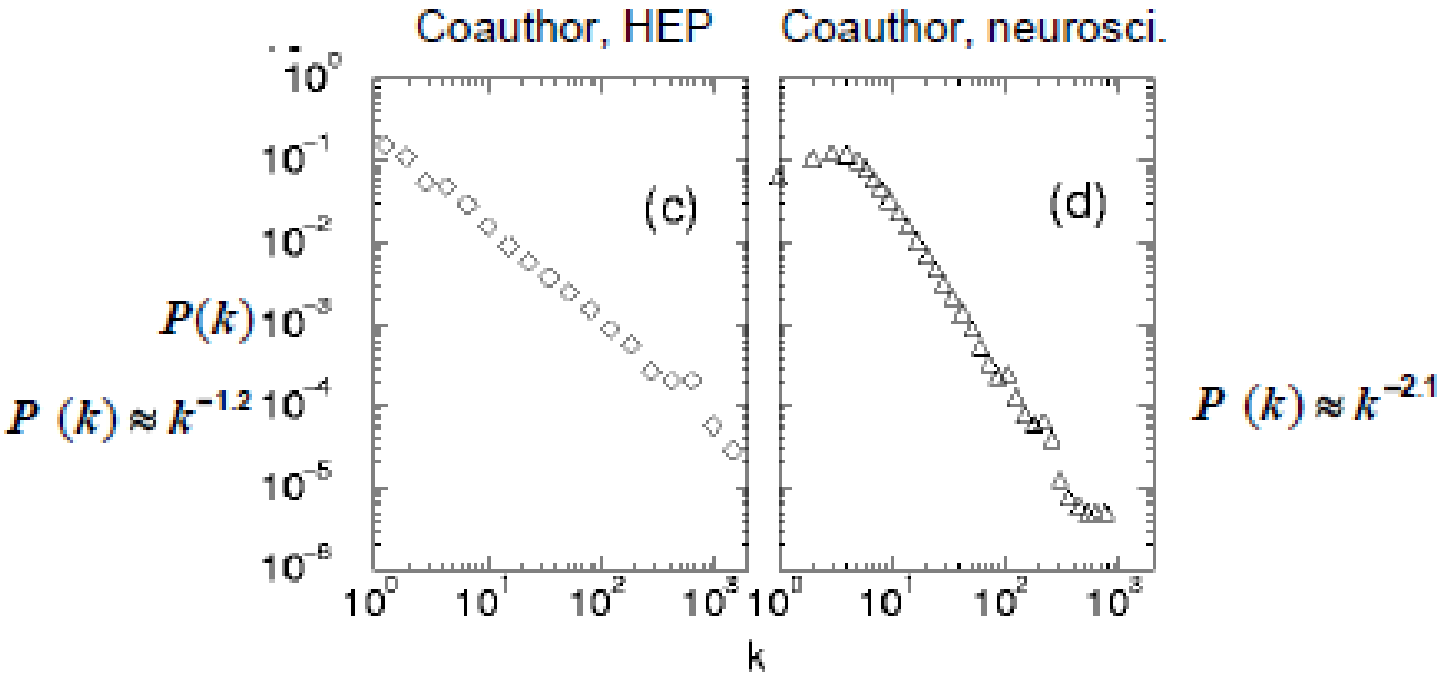


Indegree distribution

Источник: Koblenz network collection

Распределение степеней вершин: примеры

Networks of science collaborations also have power-law degree distributions



M. E. J. Newman, Phys. Rev. E 64, 016131 (2001)

A.-L. Barabási et al., cond-mat/0104162 (2001)

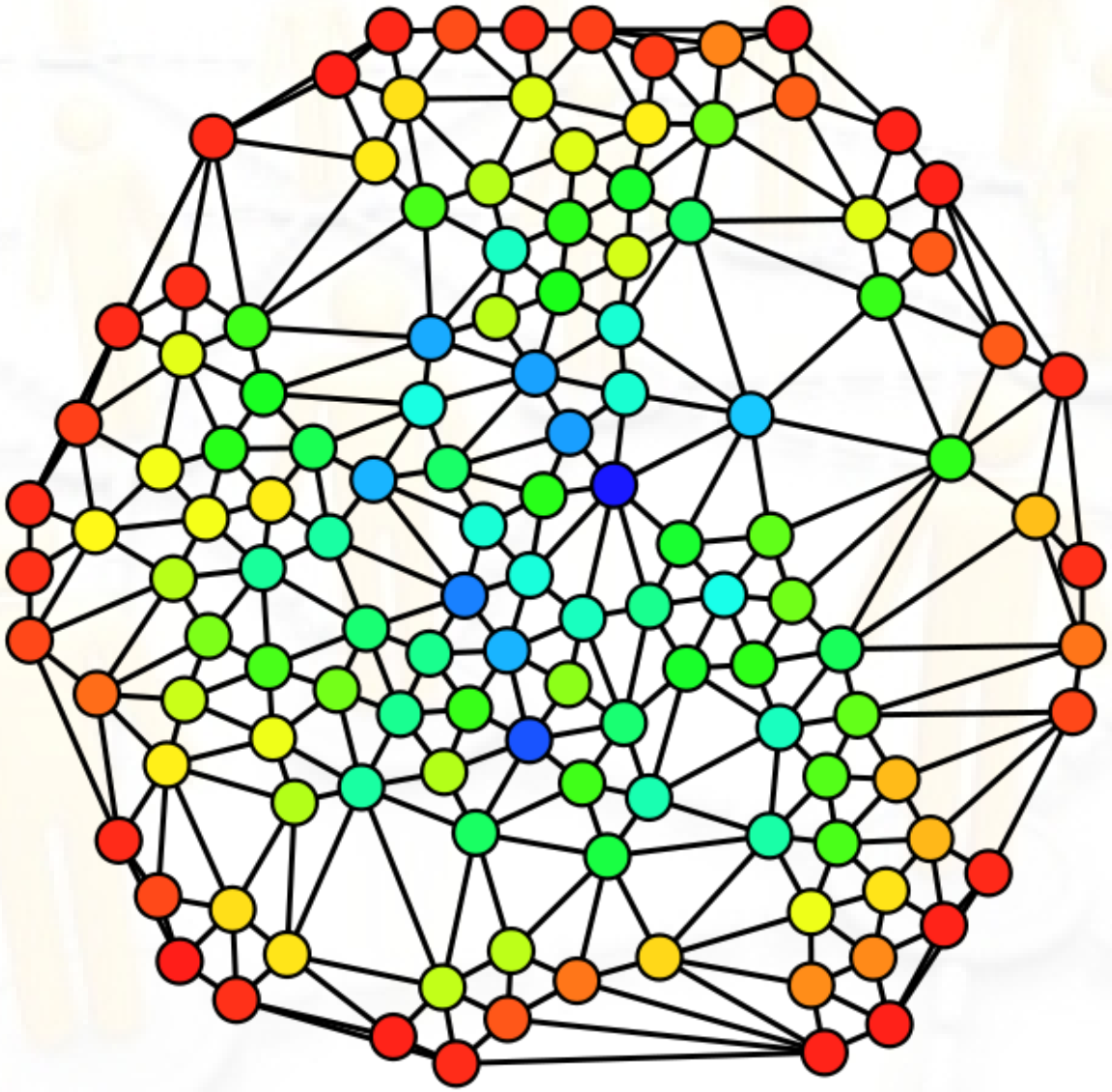
Меры центральности узлов

- *Closeness centrality* — величина, обратная к среднему расстоянию от данной вершины до произвольной
  - Другой вариант — $CC(v) = \sum_{w \in V \setminus \{v\}} 2^{-\operatorname{dist}(v,w)}$
- *Betweenness centrality* — величина, отражающая важность узла в сети:

$$BC(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

  где σ_{st} — число кратчайших путей из s в t, $\sigma_{st}(v)$ — число тех из них, что проходят через v.
- В безмасштабных сетях распределение по betweenness centrality тоже степенное.

Betweenness centrality: пример



Betweenness and closeness centrality: пример

Closeness centrality обозначена размером, Betwenness centrality — цветом

Источник: Mark Daly and Paul Rosenfeld, University of Mariland

Betweenness centrality: распределение

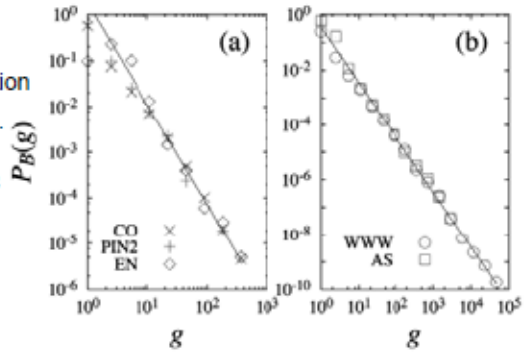
Distribution of betweenness centrality

Coauthorship

Protein interaction

Metabolic netw.

$P_B(g) \approx g^{-2.2}$



World-wide Web

Internet (AS level)

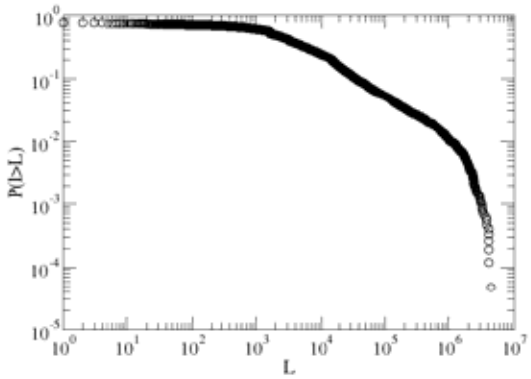
$P_B(g) \approx g^{-2}$

K. I. Goh et al., PNAS 99, 12583 (2002)

Betweenness centrality (load) distribution of the power grid

$P(l > L) \approx (2500 + L)^{-0.7}$

Q: How does the non-cumulative distribution look like in the region where the cumulative distribution is almost horizontal?



R. Albert, I. Albert, G. L. Nakarado, Phys. Rev. E 69, 025103(R) (2004)

Плотное ядро

- Достаточно удалить из сети 10% вершин самой большой степени, чтобы она распалась на мелкие компоненты
- Средний кратчайший путь 3.5 шага делает внутри 10% наиболее активных вершин

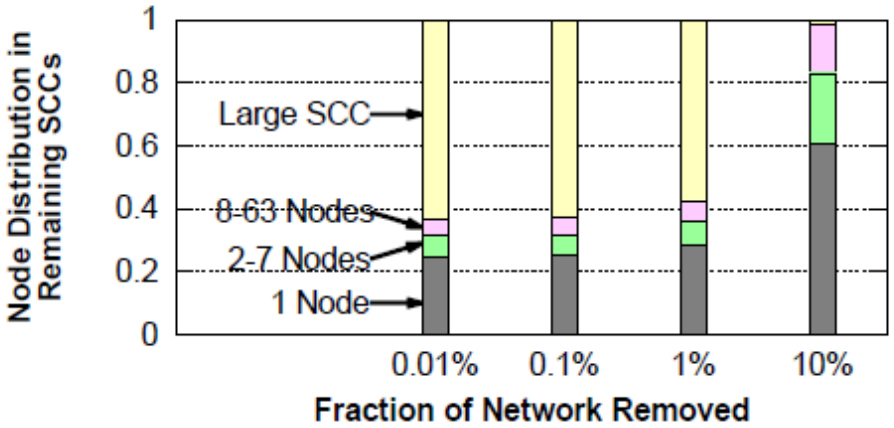


Figure 7: Breakdown of network into SCCs when high-degree nodes are removed, grouped by SCC size.

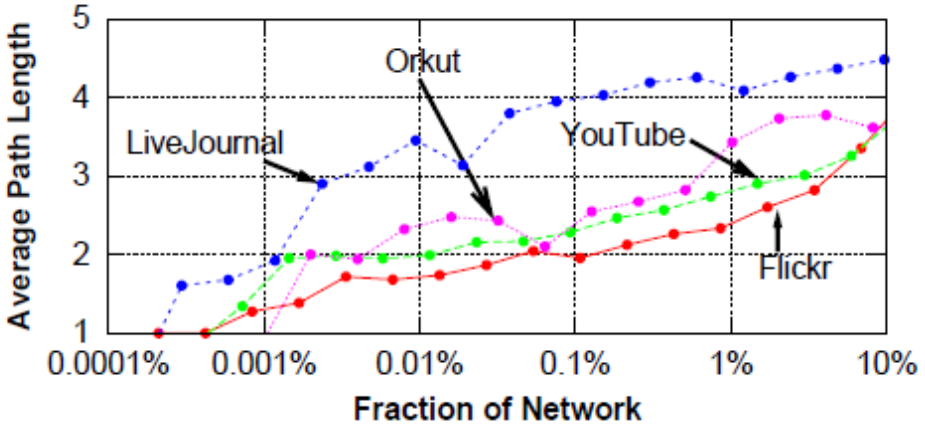


Figure 8: Average path length among the most well-connected nodes. The path length increases sub-logarithmically.

Источник: Mislove et al, 2006

Коэффициент кластеризации

- Наблюдение: знакомые одного и того же человека с высокой вероятностью знакомы между собой
- Наблюдение: знакомые одного человека разбиваются на группы знакомых между собой
- *Локальный коэффициент кластеризации*

$$C_u = \frac{\#\{(v,w) \in E \mid (u,v) \in E \,\&\, (u,w) \in E\}}{\#\{(v,w) \mid (u,v) \in E \,\&\, (u,w) \in E\}}$$

- *Средний коэффициент кластеризации*

$$\overline{C} = \frac{1}{|V|} \sum_u C_u$$

- *Глобальный коэффициент кластеризации*

$$C = \frac{3N_\triangle}{N_\wedge} = \frac{\#\{(u,v,w) \mid (u,v) \in E, (u,w) \in E, (v,w) \in E\}}{\#\{(u,v,w) \mid (u,v) \in E, (v,w) \in E\}}$$

Коэффициент кластеризации: эмпирика

Network	C	Ratio to Random Graphs Erdös-Rényi	Power-Law
Web [2]	0.081	7.71	-
Flickr	0.313	47,200	25.2
LiveJournal	0.330	119,000	17.8
Orkut	0.171	7,240	5.27
YouTube	0.136	36,900	69.4

Table 4: The observed clustering coefficient, and ratio to random Erdös-Réyni graphs as well as random power-law graphs.

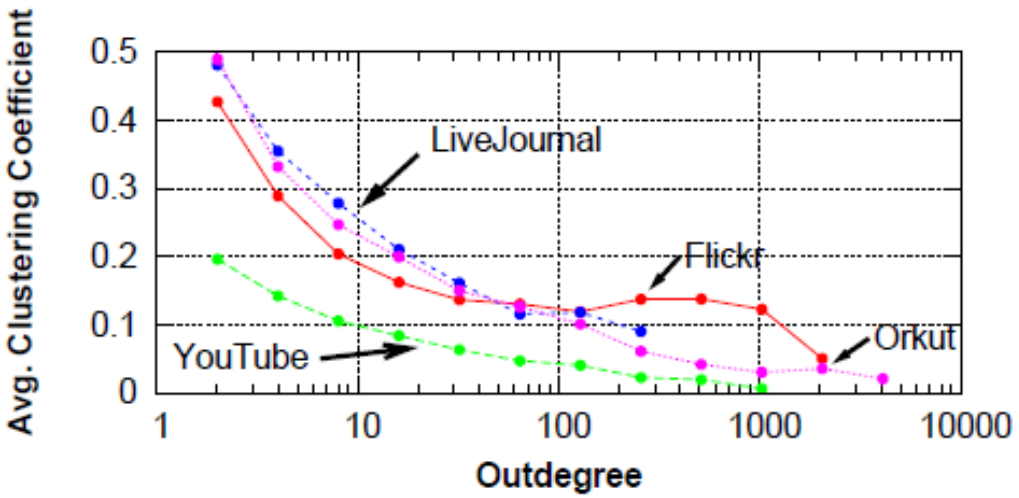


Figure 9: Clustering coefficient of users with different outdegrees. The users with few "friends" are tightly clustered.

Источник: Mislove et al, 2006

Ассортативность

- Ассортативность в широком смысле — образование связей между узлами, в чём-то схожими друг с другом (Rich men's club)
- Ассортативность в узком смысле — вершины большой степени склонны к образованию связей между собой (Клуб тысячников ЖЖ)
- Коэффициент ассортативности — коэффициент корреляции Пирсона между степенями связанных вершин:

$$r = \frac{L\sum_{i=1}^{L} d_i f_i - \left(\sum_{i=1}^{L} d_i\right)\left(\sum_{i=1}^{L} f_i\right)}{\sqrt{\left(L\sum_{i=1}^{L} d_i^2 - \left(\sum_{i=1}^{L} d_i\right)^2\right)\left(L\sum_{i=1}^{L} f_i^2 - \left(\sum_{i=1}^{L} f_i\right)^2\right)}},$$

где L — число ориентированных рёбер, d_i и f_i — остаточные степени начала и конца ребра.

Ассортативность: эмпирика

Таблица 2. Ассортативность социальных, технологических и биологических сетей. Reprinted with permission from Newman M. E. J. Mixing patterns in networks // Phys. Rev. E 67, 026126. © 2003 by the American Physical Society

	Сеть	Тип	Размер n	Ассортативность r
Социальные	соавторов по физике	неориентированная	52 909	0.363
	соавторов по биологии	неориентированная	1 520 251	0.127
	соавторов по математике	неориентированная	253 339	0.120
	сотрудничества актеров кино	неориентированная	449 913	0.208
	директоров компаний	неориентированная	7 673	0.276
	связей студентов	неориентированная	573	–0.029
	адресов электронной почты	ориентированная	16 881	0.092
Технологические	сеть электростанций	неориентированная	4 941	–0.003
	Интернет	неориентированная	10 697	–0.189
	«Всемирная паутина» (WWW)	ориентированная	269 504	–0.067
	взаимозависимости программного обеспечения	ориентированная	3 162	–0.016
Биологические	взаимодействий белков	неориентированная	2 115	–0.156
	метаболическая сеть	неориентированная	765	–0.240
	нейронная сеть	ориентированная	307	–0.226
	морская пищевая сеть	ориентированная	134	–0.263
	пресноводная пищевая сеть	ориентированная	92	–0.326

Особенности социальных сетей

Итого, социальные сети отличаются:

- Маленьким диаметром и средним расстоянием между вершинами;
- Степенным законом распределения степеней вершин и betweenness centrality;
- Высоким коэффициентом кластеризации;
- Ассортативностью;
- Наличием тесно связанного ядра.

Особенности социальных сетей

Итого, социальные сети отличаются:

- Маленьким диаметром и средним расстоянием между вершинами;
- Степенным законом распределения степеней вершин и betweenness centrality;
- Высоким коэффициентом кластеризации;
- Ассортативностью;
- Наличием тесно связанного ядра.

Конечная цель моделирования: построить модель, в которой будут отражены все эти свойства.

Модели случайных графов

- Модель Эрдеша–Реньи (1959): каждое ребро возникает с вероятностью p независимо от других.
  - \+ Маленькое среднее расстояние
- Модель Уоттса–Строгатца (1998): берётся кольцо, каждая из вершин которого соединена с K соседними. После этого каждое ребро с вероятностью β «переключается» на случайную вершину.
  - \+ Маленькое среднее расстояние и большой коэффициент кластеризации
  - – Нереалистичное распределение степеней вершин
- Модели с предпочтительным присоединением (preferential attachment): Боллобаш–Риордан, Барабаши–Альберт, Бакли–Остхус и др. Идея: постепенный рост сети, новые вершины с большей вероятностью присоединяются к тем, у кого текущая степень выше.
  - \+ Парето-распределение степеней, маленький диаметр, положительный коэффициент ассортативности
  - – Маленький коэффициент кластеризации

Эпидемические модели в случайных графах

Grant Schoenebeck, 2010:

- Все изученные сети удовлетворяют степенному закону
- Но целиком мы можем анализировать только онлайн-сети
- Что, если все онлайн-сети формируются по одному и тому же закону?
- Идея: онлайн-сеть как результат случайного распространения эпидемии по обычной сети
- **Результат**: при распространении эпидемии по сети Уоттса–Строгатца может получиться сеть, удовлетворяющая степенному закону

Эпидемическая модель: результат моделирования

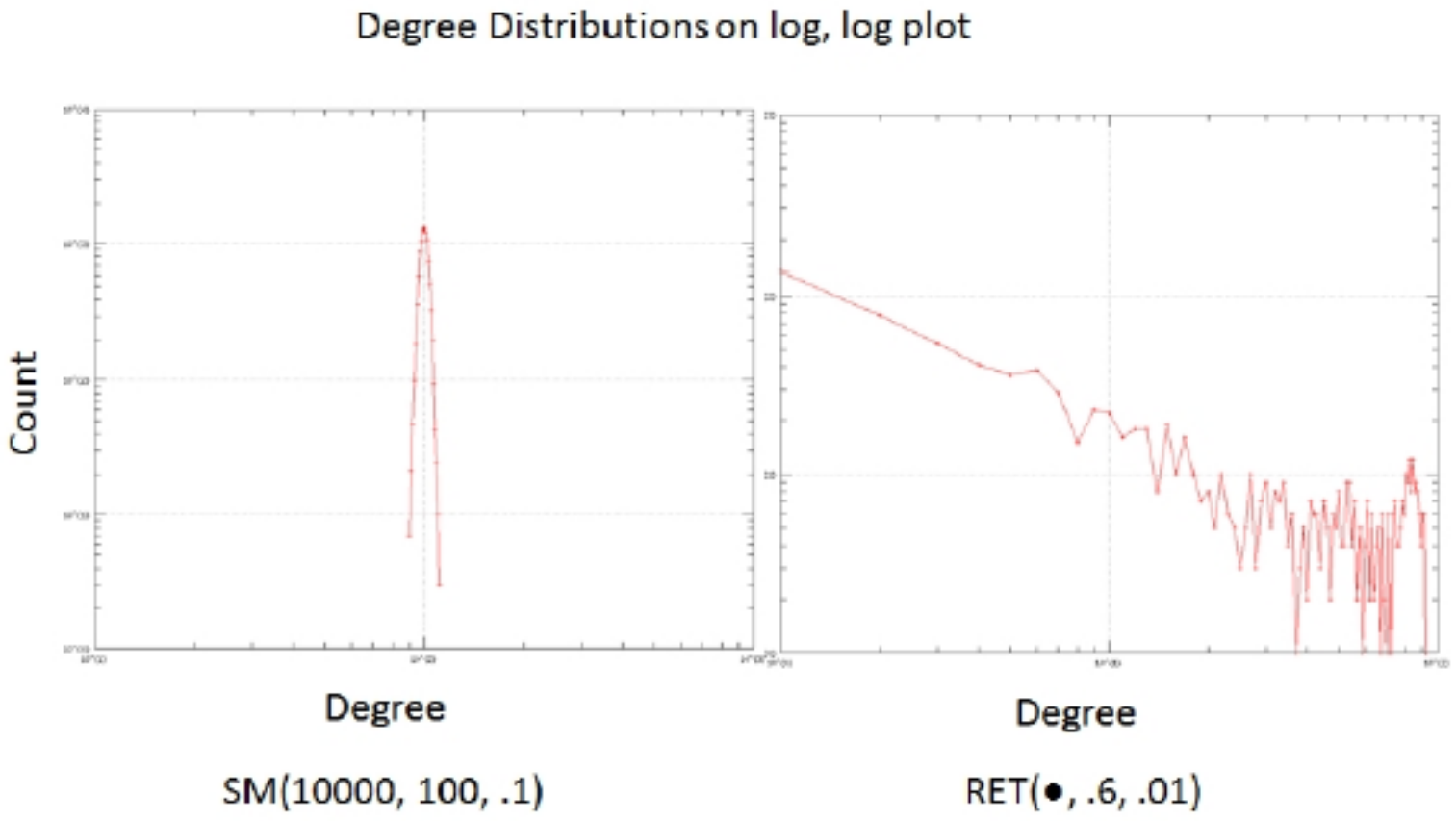


Figure 1: A degree plot of $WS(1000, 100, .1)$ and $RET_{WS(10000,100,.1)}(1000, .6, .01)$

Кооперативные теоретико-игровые модели

- Структура сети может определять полезности агентов, причём имеют место экстерналии
- Например, может быть важен состав второго круга (сети соавторства, военные союзы)
- *Модель Джексона-Волински*: каждой конфигурации сети сопоставлены выигрыши агентов.
- Установление или разрыв связи бесплатны, но для установления связи необходимо согласие обоих агентов, а разорвать связь можно единолично
- Ищутся устойчивые конфигурации
- Устойчивые конфигурации могут быть неэффективными даже при трансферабельной полезности

Модель Джексона–Волински, пример 1

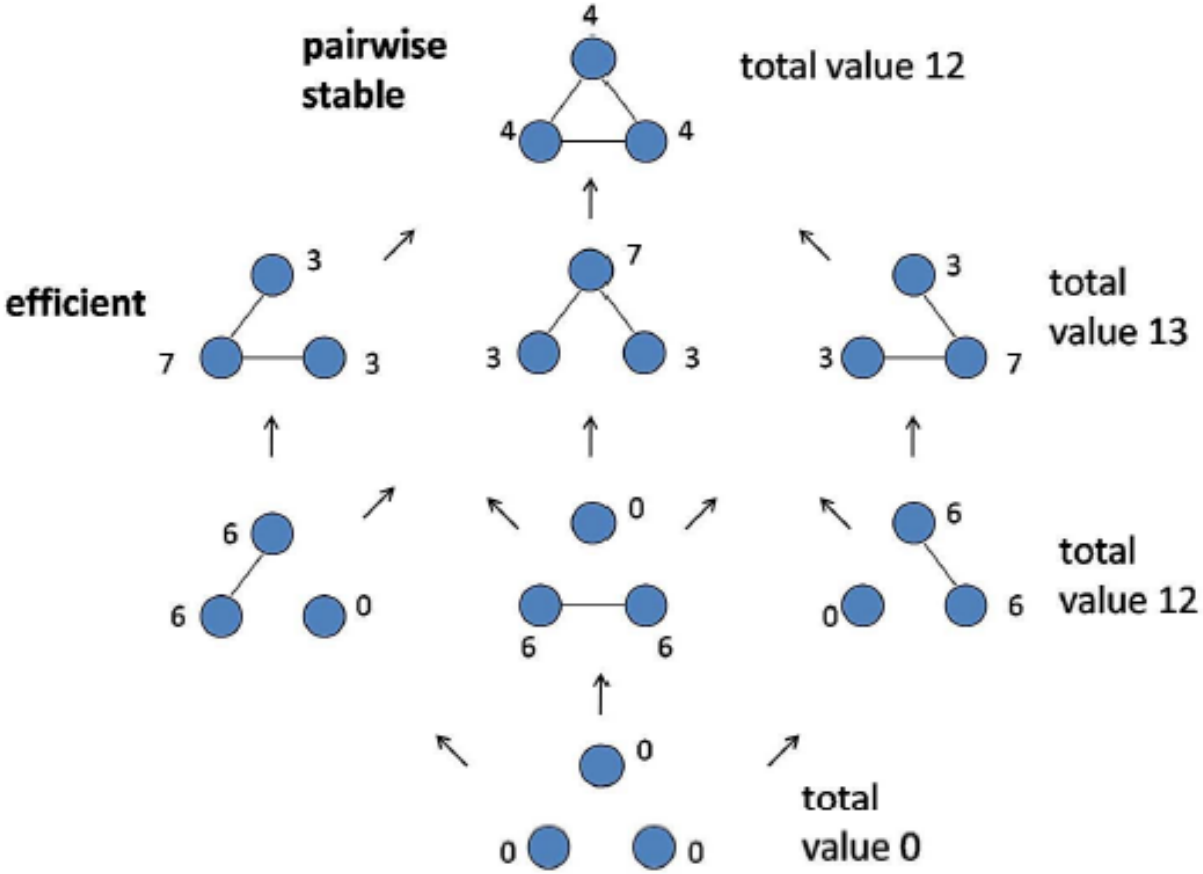


Figure 4.3. Payoffs to each node as a function of the network. Two-link networks result in the maximal overall payoffs. The arrows indicate changes networks that benefit both nodes associated with adding a link or at least one node who can delete a link. The unique pairwise stable network in this simple example is the complete network.

Источник: M.O.Jackson, “An Overview of Social Networks and Economic Applications”

Модель Джексона–Волински, пример 2

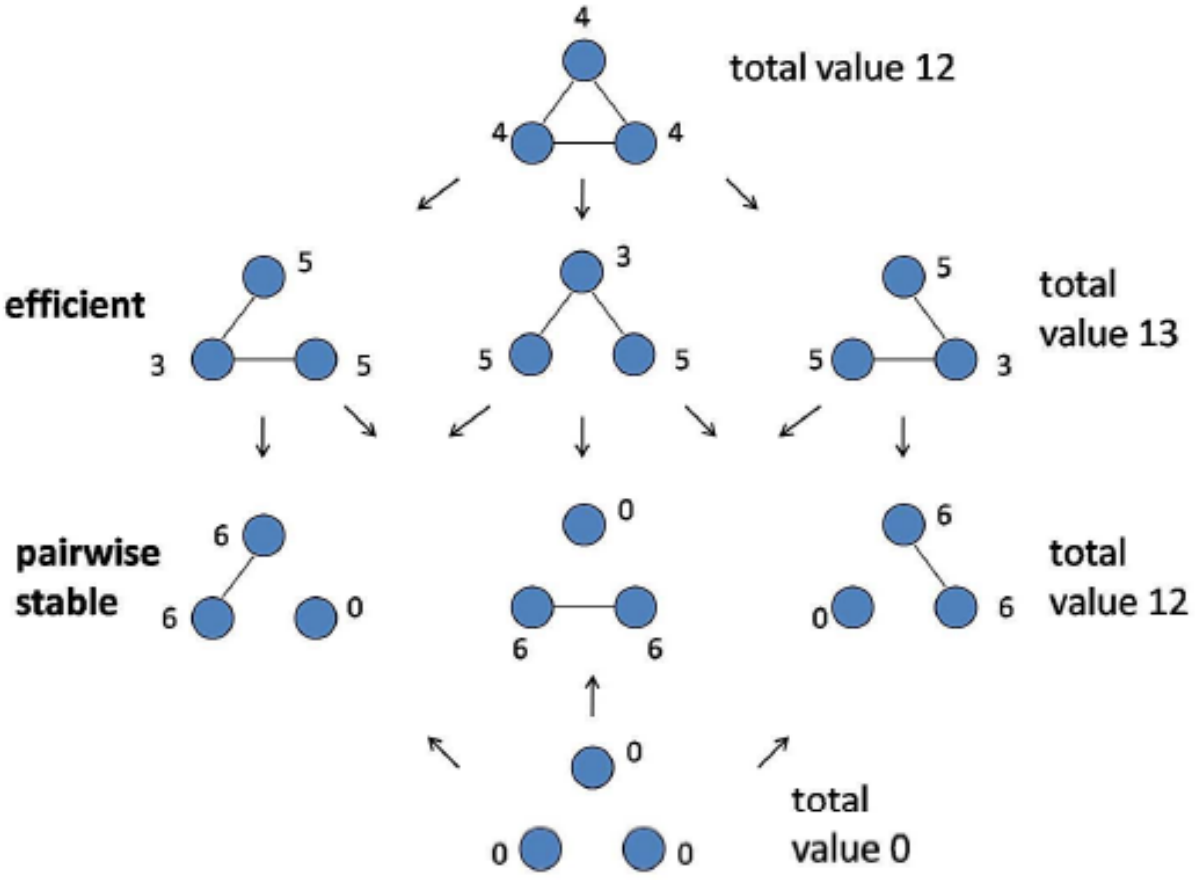


Figure 4.3. Payoffs to each node as a function of the network. Two-link networks result in the maximal overall payoffs. The arrows indicate changes in the network that benefit both nodes associated with adding a link, or one of the nodes involved in deleting a link. The pairwise stable networks are the one-link networks.

Источник: M.O.Jackson, “An Overview of Social Networks and Economic Applications”

Модель Джексона–Волински, пример 3

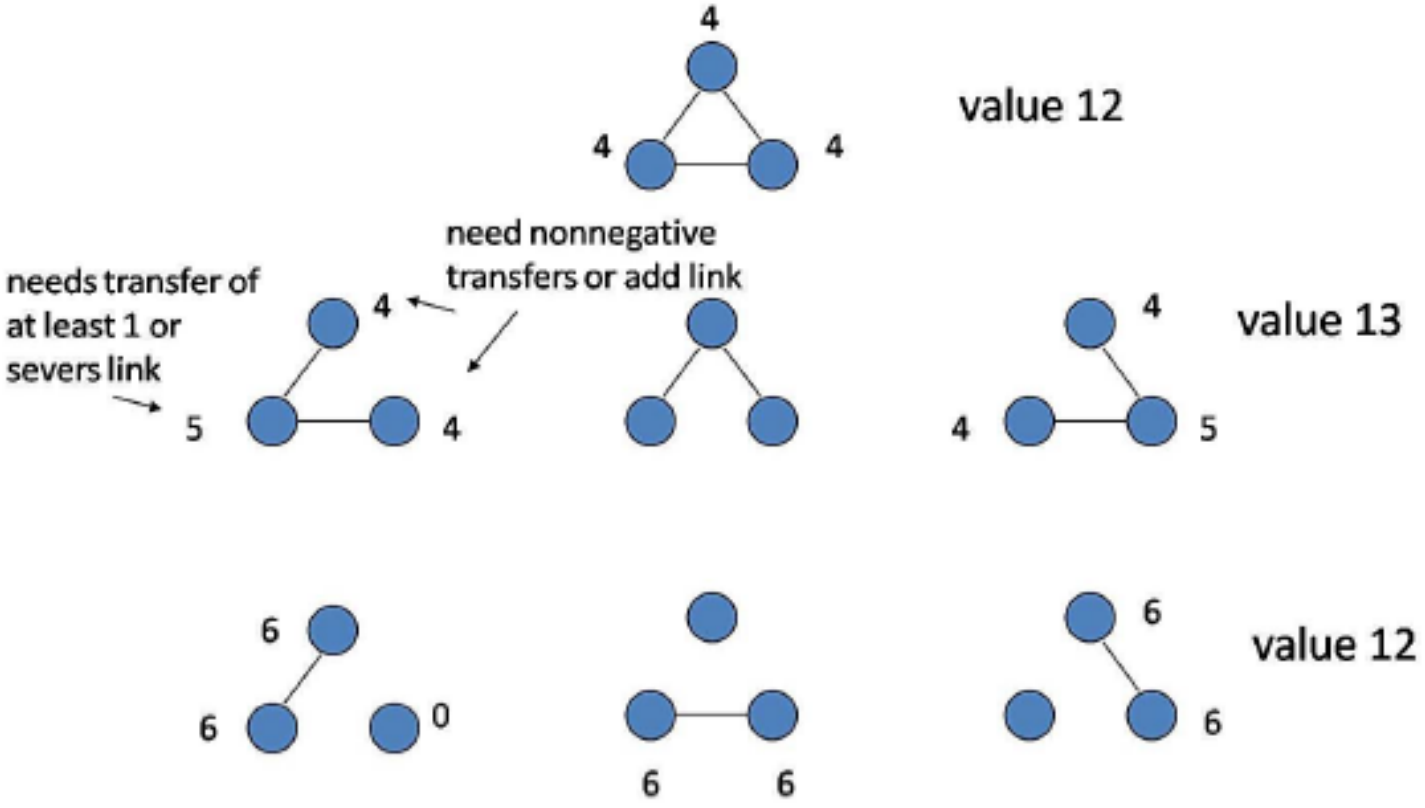
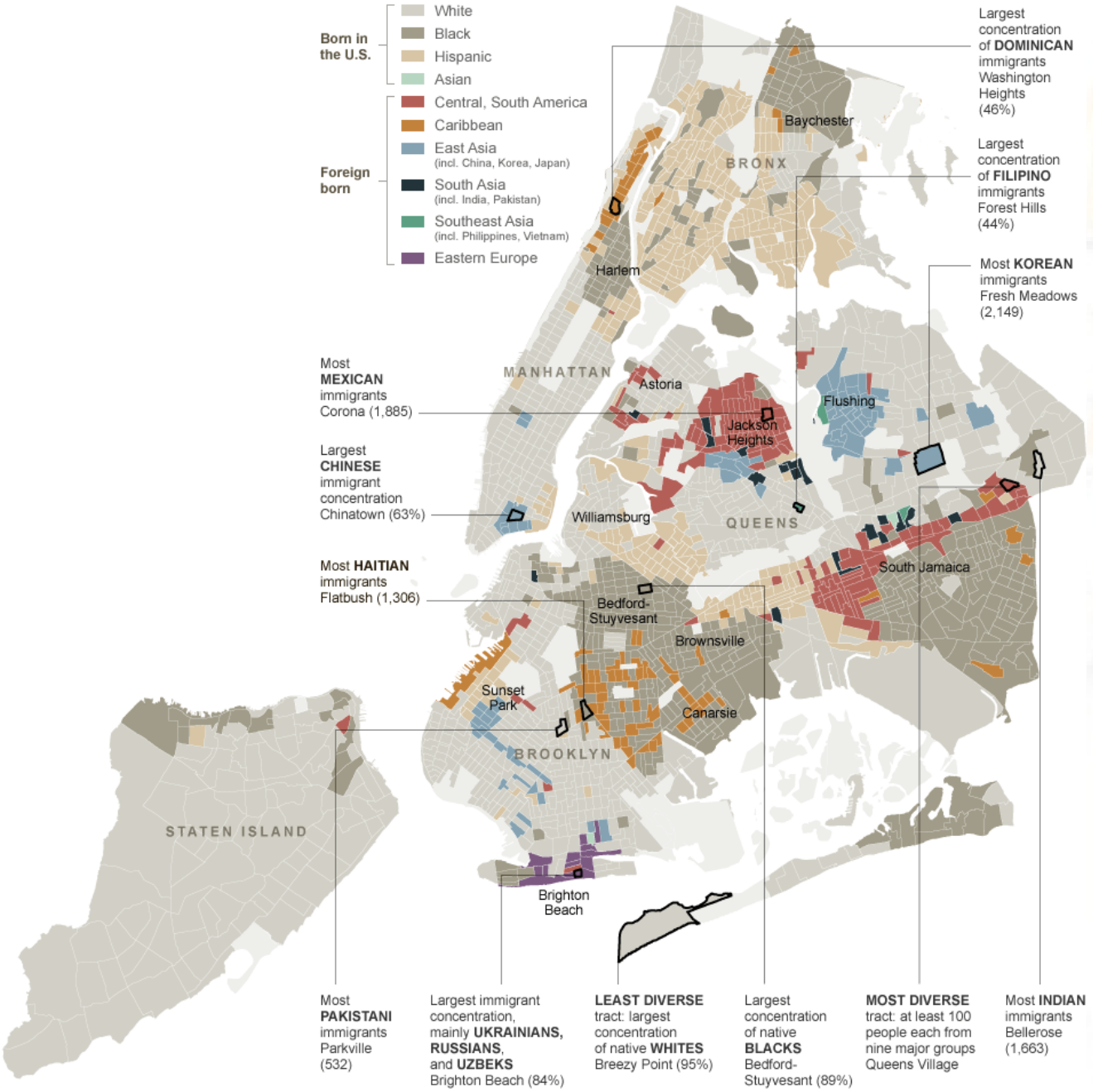


Figure 4.3. The impossibility of maintaining the total utility maximizing network as being pairwise stable, regardless of transfers between the peripheral nodes and center node.

Источник: M.O.Jackson, “An Overview of Social Networks and Economic Applications”

Модель Шеллинга: случайность + стратегия

- Агенты разделены на две группы, члены каждой из которых предпочитают находиться рядом с себе подобными
- Изначально агенты распределяются случайно по клеткам квадратной доски
- Агент доволен, если на соседних с его клетках живут хотя бы два агента его группы
- Выбирется недовольный агент и переселяется в ближайшую клетку, где он будет доволен
- И так далее
- В итоге могут возникнуть кластеры, даже если изначально всё было хорошо перемешано.



Источник: The New York Times

Модель Чайес-Боргса

Borgs, Chayes, Ding, Lucier, “The Hitchhiker’s Guide to Affiliation Networks: A Game-Theoretic approach”, 2010

- Первая полностью некооперативная теоретико-игровая модель
- Также отражает принцип вовлечённости каждого человека во множество разных групп
- Кратко: каждый агент может организовать собрание, на которое позвать любое множество агентов. У собрания есть интенсивность, которой пропорциональна стоимость организации. Связь образуется, если агенты побывали вместе на собраниях достаточно большой суммарной интенсивности. Полезность — число связей.

Модель Чайес-Боргса: стратегии и издержки

- V — множество агентов
- Каждый агент может образовать произвольный набор собраний $P \subset V$ с интенсивностями $r > 0$
- Фиксированные издержки от собрания интенсивности r равны rb
- Предельные издержки на каждого дополнительного участника равны rc
- Итого: $C(P_{v,i}) = r_{v,i}(c|P_{v,i} \setminus \{v\}| + b)$, где v — агент, а i — номер собрания.

Модель Чайес-Боргса: формирование сети и полезность

- Связь между u и v образуется, если суммарная интенсивность собраний с участием обоих превыщает 1
- Неважно, кто организовал собрания: u, v или третья сторона
- Каждая связь приносит a обоим агентам
- Формально: $(vu) \in E$, если $\sum_{(w,i)\colon u,v \in P_{w,i}} r_{w,i} \geq 1$
- $N_v = \#\{u\colon (vu) \in E\}$
- $U_v = aN_v - \sum_i C(P_{v,i})$

Модель Чайес-Боргса: равновесия

Обозначение: $\gamma = \frac{a}{c}$

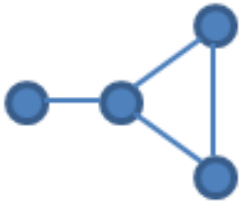
 $\gamma > 1/4$	 $\gamma > 1/3$	 $1/3 < \gamma < 1$
 $1/2 \leqslant \gamma < 1$	 $1/2 \leqslant \gamma < 1$	 Not supportable

Table 1: Sample connection graphs, with the range of parameter γ in which they are supportable as strong subgraphs.

Модель Чайес-Боргса: равновесия

- Конфигурация «Звезда» не поддерживается в равновесии
- Коэффициент кластеризации не меньше, чем $\frac{1}{\overline{d}_G}$
- Многие распределения степеней вершин могут быть поддержаны в равновесии с ограниченным размером собраний

- Теорема (Измалков)

  Если хотя бы одна конфигурация поддерживается в равновесии, то любая конфигурация поддерживается.

Предложение: дифференцированные издержки

- Предположим, что между любыми двумя агентами u и v есть «расстояние» $\rho_{uv} \geq 0$, такое что $\rho_{uu} = 0$
- Расстояние не обязательно является метрикой и даже не обязательно симметрично
- ρ_{uv} толкуется как стоимость приглашения v для агента u
- $C_v(P) = r(c \sum_{v \in P} \rho_{uv} + b)$
- Исходная формализация является частным случаем для $\rho_{uv} = \delta_{uv}$

Предложение: дифференцированные выигрыши

- Предположим, что для любых двух агентов u и v задан «потенциал» $\pi_{uv} \geq 0$, такой что $\pi_{uu} = 0$
- Потенциал не обязательно симметричен
- π_{uv} толкуется как выигрыш от контакта с v для агента u.
- $U_v = a \sum_{(uv) \in E} \pi_{uv} - \sum C_v(P)$.
- Исходная формализация является частным случаем для $\pi_{uv} = \delta_{uv}$.

Спасибо!

mailto:musatych@gmail.com
http://musatych.livejournal.com