Введение в эконометрику сетей

Владислав Морозов

27 декабря 2016 г.

1 Определения, нотация и мотивация

2 Примеры вопросов и три постановки

3 Влияние структуры

4 Идентификация DGP

5 Формирование сетей

6 Работа с выборочными данными

Основные определения

Определение

Графом называется пара (V,E) из множества вершин V и множества ребер E.

Граф представим в виде матрицы смежности $\boldsymbol{A}$:

$$A_{ij} = \begin{cases} 1, & ij \in E \\ 0, & ij \notin E \end{cases} \tag{1}$$

Плотность и распределение степеней

Определение

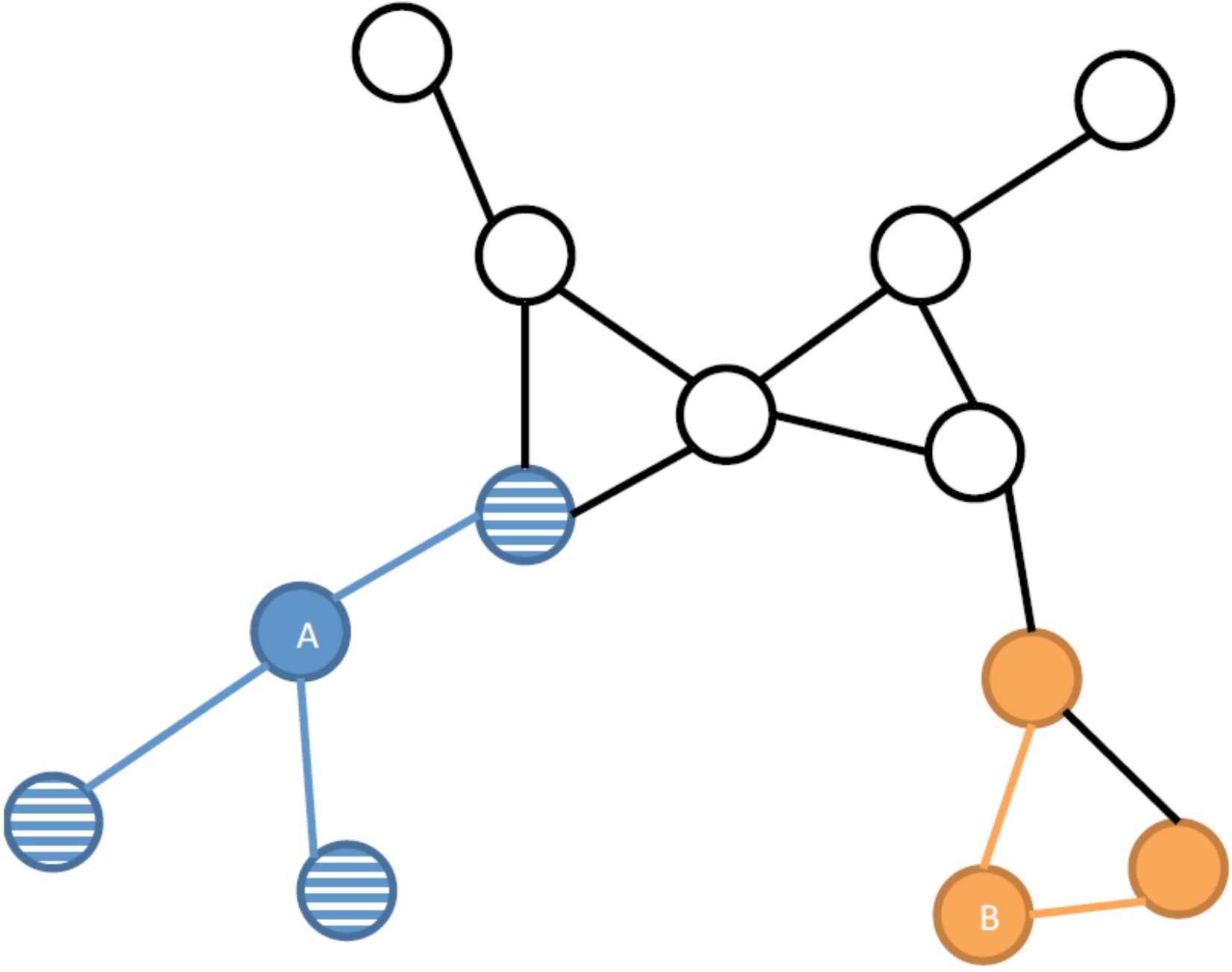
Плотность: сколько ребер из возможных C_n^2 *существует:*

$$density = \frac{1}{\mathrm{C}_n^2} \sum_{i<j} A_{ij} \tag{2}$$

Определение

Распределение степеней – CDF:

$$F_d(x) = \frac{|\{i : d_i \leqslant x\}|}{n} \tag{3}$$


A
B

Кластеринг и гомофилия

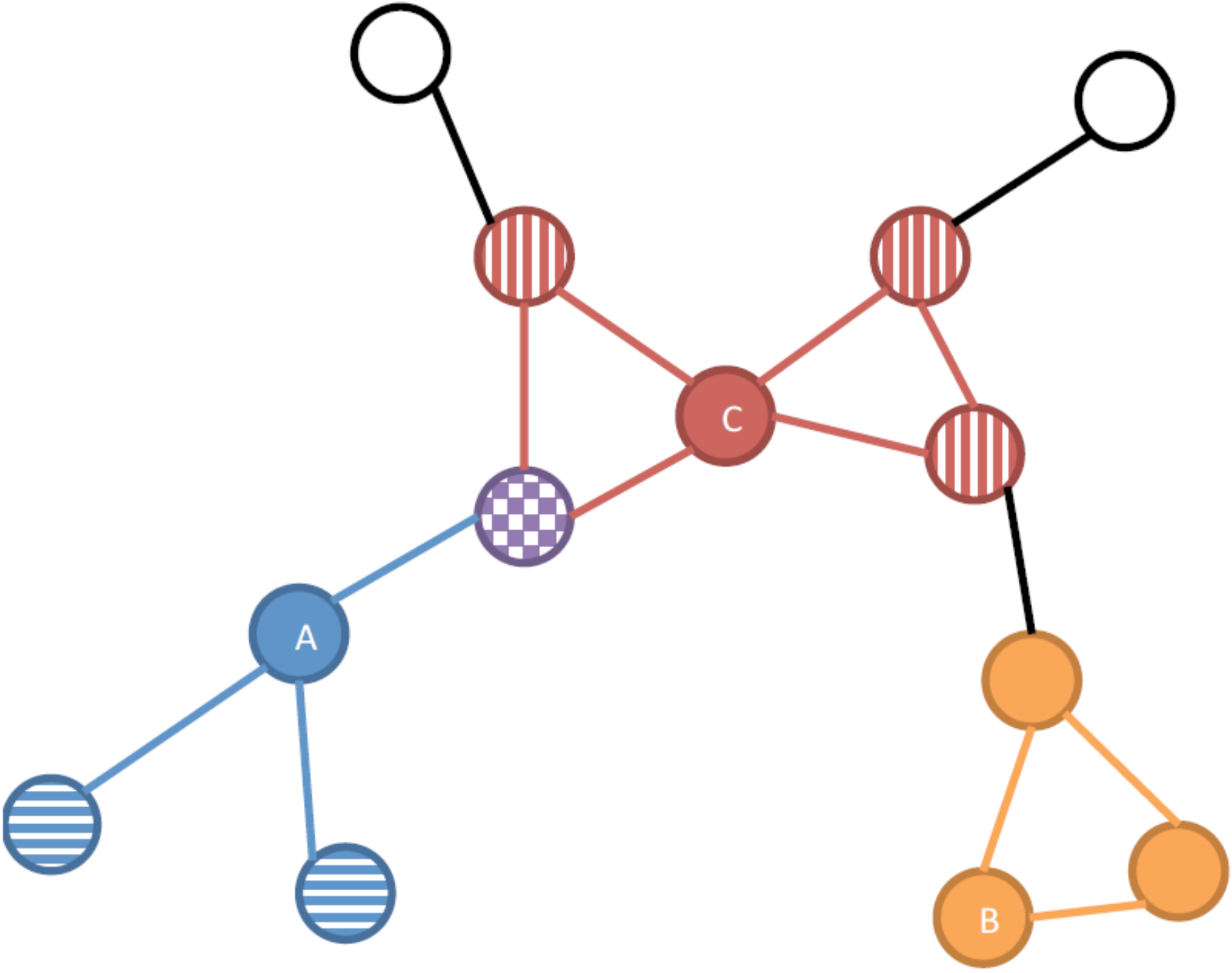
Определение

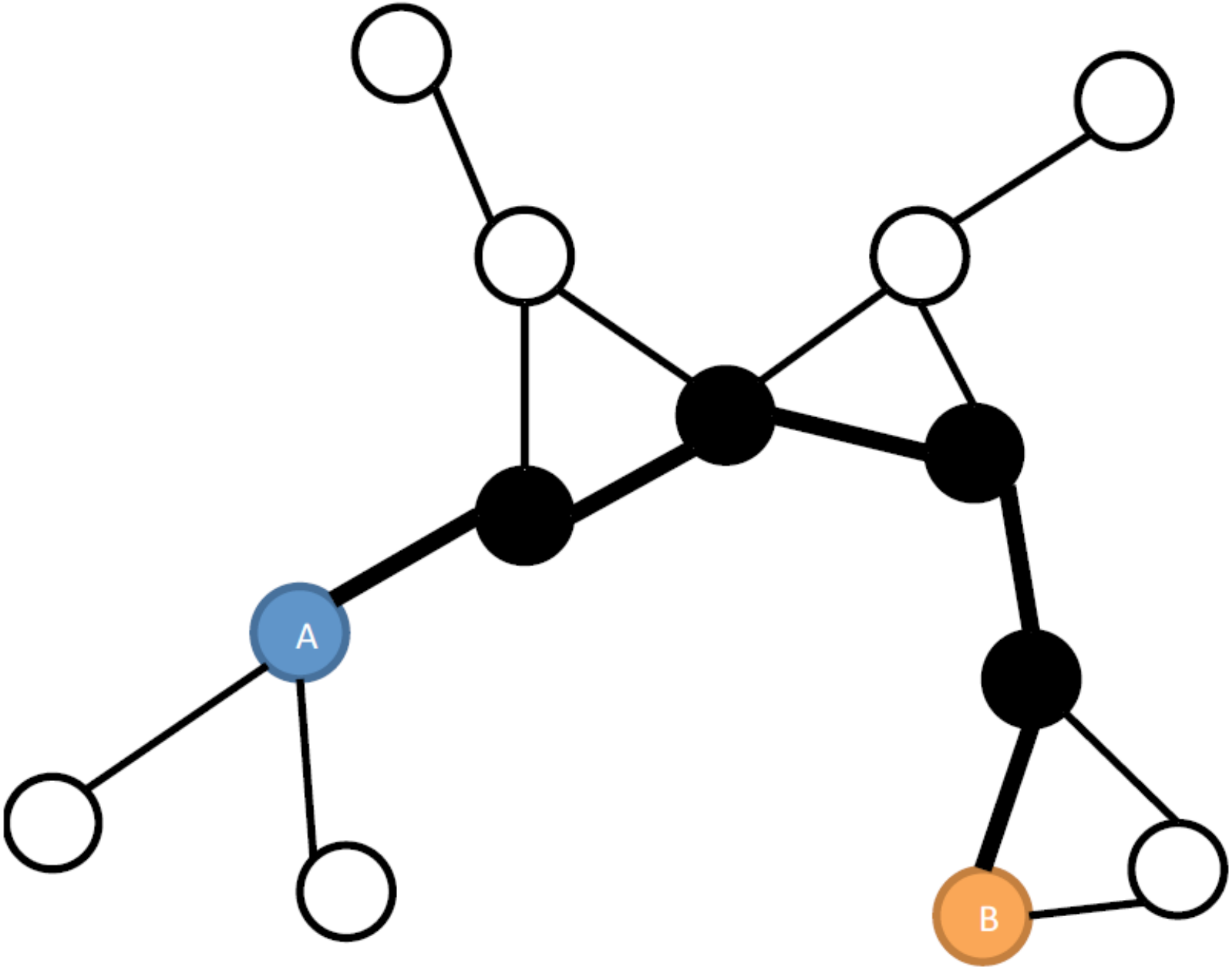
Локальный коэффициент кластеринга показывает, какова вероятность, что два соседа вершины i – j и k связаны сами:

$$c(g) = \frac{\sum_{j<k} A_{ij} A_{ij} A_{jk}}{\mathrm{C}_{d_i}^2} \tag{4}$$

Гомофилия: подобное притягивает подобное.

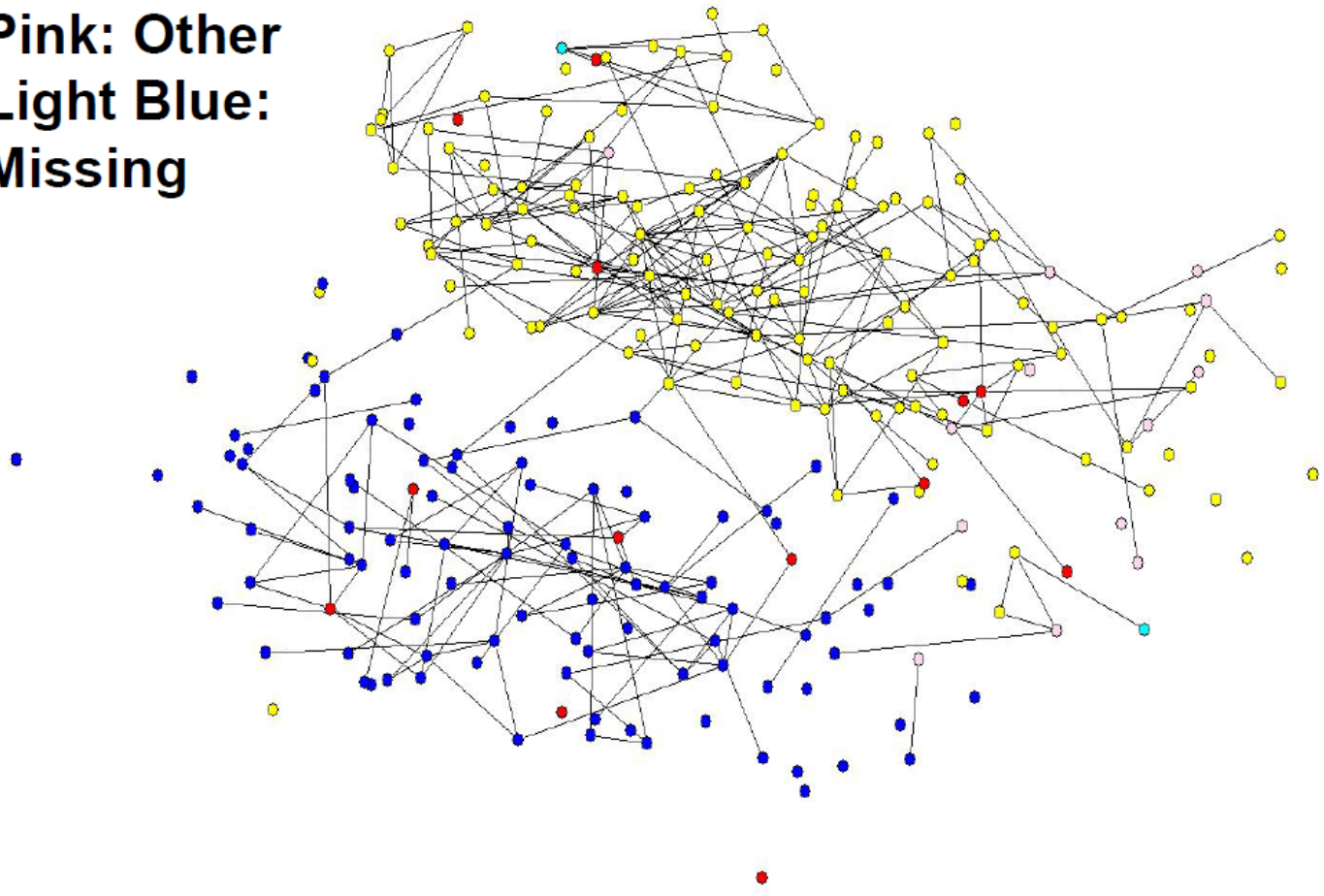
- Корреляция в ребрах из-за каких-то личных характеристик
- Не путать со структурной транзитивностью (Graham, 2014): корреляция из-за стимулов к образованию связей


C
A
B


A
B

Blue: Black
Reds: Hispanic
Yellow: White
Pink: Other
Light Blue:
Missing

"strong friendships"
cross group links less than half as frequent



Data Source	Average Degree	Density	Clustering	Average Path Length	Number of Networks	Average Number of Nodes
AddHealth	3.361	0.042	0.411	--	14	80.36
China Villages	3.266	0.113	--	2.578	185	28.82
Harvard Dorm	10.830	0.019	0.181	3.260	1	569
Harvard Facebook-Based	7.917	0.003	0.174	4.571	1	2360
Karnataka Villages	17.378	0.089	0.303	2.337	75	213.37
Malawi Villages	6.189	0.061	0.234	3.089	21	133.48
Uganda Village	8.300	0.069	0.230	2.500	1	119

Эндогенные и экзогенные эффекты

Manski (1993) – два канала социального влияния:

- *Эндогенные* эффекты: влияние исходов других вершин на исход вершины i. Работают как каналы связи + эффект мультипликатора.
  Пример: успеваемость i меняется вместе с успеваемостью его окружения
- *Экзогенные* (*контекстуальные*) эффекты: влияние характеристик вершины i на ее исходы.
  Пример: на успеваемость влияет социоэкономический статус вершины

Проблема: эффекты сложно вычленяются. *Пример*: успеваемость зависит от учителя

Какие вопросы могут нас интересовать?

- Как жители деревни в Танзании организуют risk-sharing? Какие есть внешние эффекты при этом? (Comola, 2016)
- Как распространяется знание о микрофинансировании? (BDSJ, 2013)
- Как студенты заводят дружбу? (Польдин, Креховец 2014)
- Насколько сильна гомофилия? (Hsieh, Lee 2015)

Три типа вопросов

Вопрос

Как сетевая структура влияет на социальные исходы?

Вопрос

Какой DGP лежит в основе данных?

Вопрос

Как формируются сети?

Влияние характеристик сети на исходы

Простой случай:

$$y_i = \alpha + \beta\omega(G) + \varepsilon_i \tag{5}$$

Можно оценивать:

- Network-wide регрессии, если у нас много сетей и есть какие-то общие исходы (распространение информации о МФ у BDSJ (2013))
- Node-level регрессии: влияние индивидуальных сетевых характеристик на какие-то его исходы

Классическая линейная модель Манского

Manski (1993) предложил классическую линейную модель влияния сетевой структуры на исходы в обществе из N индивидов:

$$y_i = \alpha + \beta \sum_{j=1}^{N} A_{ij} y_j + \boldsymbol{\eta} \boldsymbol{x}_i + \boldsymbol{\gamma} \sum_{j=1}^{N} A_{ij} \boldsymbol{x}_i + \varepsilon_i \tag{6}$$

С предпосылкой, что $\mathrm{E}(\varepsilon_i | \boldsymbol{x}, A) = 0$.
Если $1/\beta$ не собственное значение A, то записывается в следующей форме (Manski Reduced Form):

$$\boldsymbol{y} = \alpha(\boldsymbol{I} - \beta A)^{-1}\boldsymbol{1} + (\boldsymbol{I} - \beta A)^{-1}(\eta \boldsymbol{I} + \gamma A)\boldsymbol{x} + (I - \beta A)^{-1}\varepsilon \tag{7}$$

Можно ли оценить коэффициенты?

Теорема

Если $|\beta| < 1$, $\eta\beta + \gamma \neq 0$, $A_{ij} = (N-1)^{-1}$ для $i \neq j$, а $A_{ii} = 0$, то $(\alpha, \beta, \boldsymbol{\eta}, \boldsymbol{\gamma})$ не определяются однозначно.

Насколько эта проблема общая?

Теорема об идентификации

Bramoulle, Djebbari и Fortin (2009) дают теорему:

Теорема

Если $\eta\beta + \gamma \neq 0$ и $\boldsymbol{I}$, A и A^2 линейно независимы, то существуют оценки для $(\alpha,\beta,\boldsymbol{\eta},\boldsymbol{\gamma})$.

Насколько это жесткое условие?
Blume, Brock, Durlauf и Jayaraman (2015):

Теорема

Если $\sum_{j=1}^{N} A_{ij} = 1$ и $A_{ii} = 0$, и $\boldsymbol{I}$, A и A^2 линейно зависимы, то A – блочно-диагональная матрица с блоками одного размера N_t, а все ненулевые элементы имеют вид $(N_t - 1)^{-1}$.

А если мы не знаем A?

Manresa (2014) рассматривает регрессию (с панельными данными):

$$y_{i,t} = \alpha + \beta \sum_i \frac{A_{ij}}{d_i} x_{j,t} + \delta x_{i,t} + \varepsilon_{i,t} \tag{8}$$

В случае разреженной сети pooled LASSO (Tibshirani, 1996;Meinshausen and Yu, 2009) может получить оценки для коэффициентов и матрицы.

Если есть какая-то ненаблюдаемая гетерогенность, то требуются наблюдания в «достаточном» числе периодов. Graham (2014) выводит оценки для этого числа.

Возможные нелинейности

Два основных канала:

- Нелинейные формы влияния:

$$y_i = f\left(\sum_{j=1}^{N} A_{ij} y_j, x_i, \sum_{j=1}^{N} A_{ij} x_j\right) \tag{9}$$

- Нелинейная реакция на исходы других участников (например, $\min_{A_{ij} \neq 0} y_j$)

Оценка нелинейных форм

Теорема

Если $A_{ij} = (N-1)^{-1}$ для $i \neq j$, а $A_{ii} = 0$, то невозможно однозначно оценить f.

Blume, Brock, Durlauf, Ioannides (2011) – хороший обзор случаев, когда оценка возможна: бинарный и множественный выбор особенно хорошо оцениваются

Tincani (2015) – пример нелинейного агрегирования исходов других участников на примере успеваемости в Перу.

Модели формирования сетей

Несколько основных классов:

- *Conditional edge independence*: агенты формируют связи только на основе своих атрибутов и шоков
- *Higher order dependence*:
  - ERGM
  - SERGM
  - SUGM
  - Strong homophily

О проблемах с информацией

В оценке параметров могут возникать существенные проблемы из-за недостатка информации. Два крайних случая:

- В модели Эрдеша-Реньи все ребра генерируются независимо друг от друга. N эффективных наблюдений
- С другой стороны: все ребра могут быть полностью коррелированы: одно эффективное наблюдение

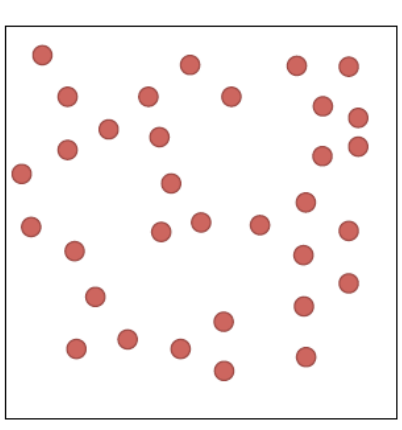
Нетривиальная проблема.

Модель Эрдеша-Реньи

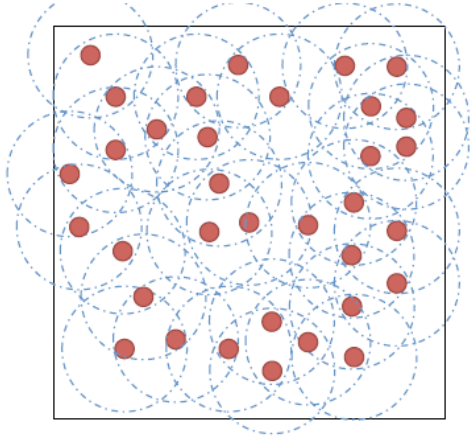
Все ребра независимо друг от друга возникают с вероятностью p.
Свойства:

- $\mathrm{E}(d_i) = (N-1)p$;
- Нет кластеринга
- Не отражает гомофилию

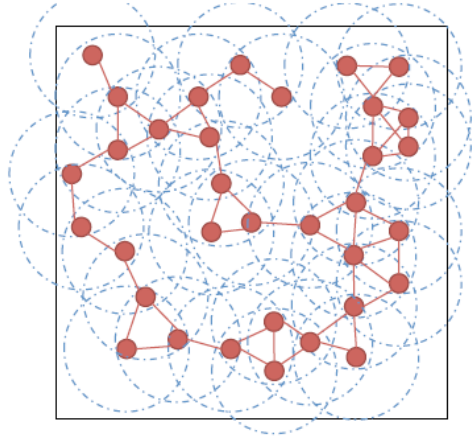
Если данные только по одной сети, мы оценивает p как число, хотя оно может быть функцией. Не можем различать между плотными и разреженными сетями.



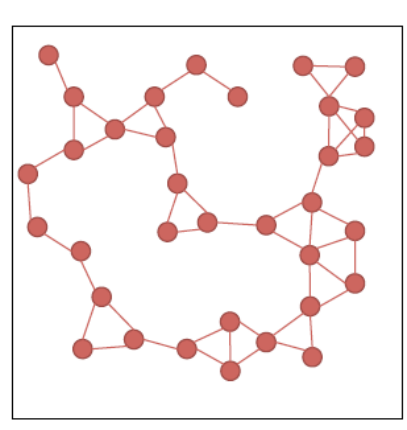
(A) Nodes



(B) Radii



(C) Links



(D) Network

Preferential Attachement и гибридные модели

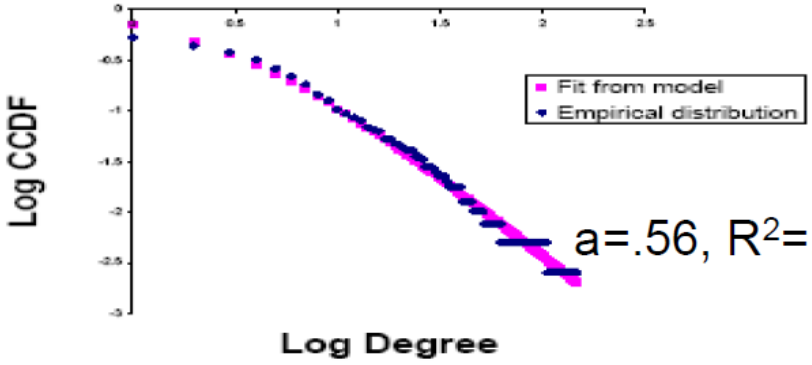
Jackson, Rogers (2007) предлагают простой в оценке гибрид модели Эрдеша-Реньи и модель с preferential attachment (Barabasi, Albert 2001):

$$F(d) = 1 - \left(\frac{m + \alpha m x}{d + \alpha m x} \right)^x \tag{10}$$

- $x = 2/(1 - \alpha)$
- m – число связей, формируемых каждый период в модели BA
- α – пропорция связей, формирумых равноверно случайно в периоде

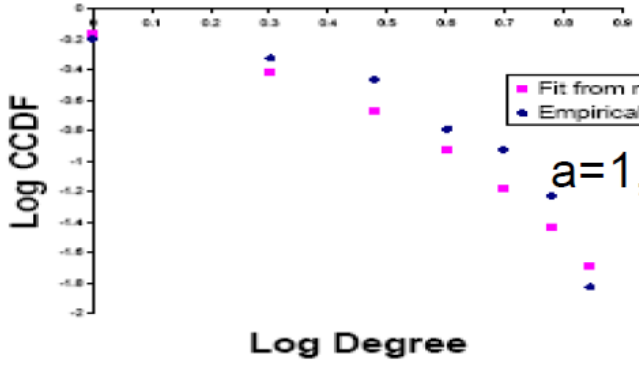
Оценивается m, потом минимизируется расстояние между фактическим распределением и теоретическим, чтобы найти α.

Small World Citations



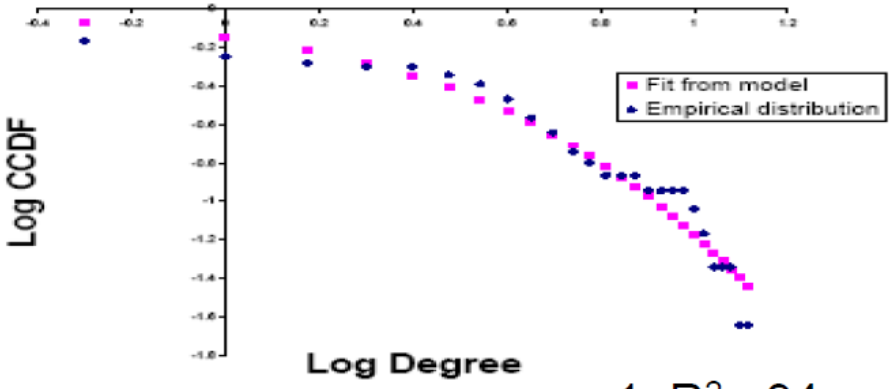
a=.56, R^2=.98

Prison Inmate Friendships



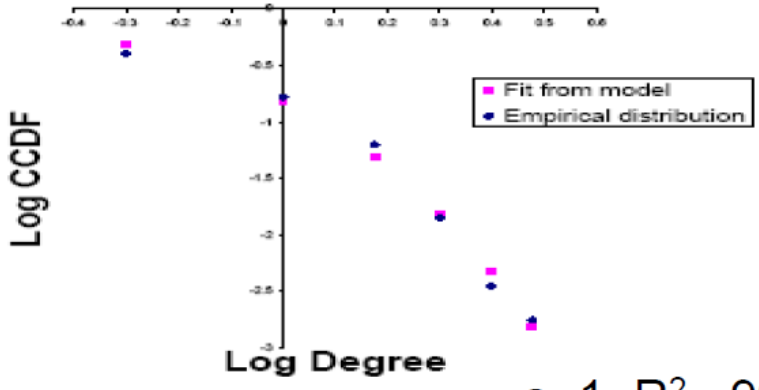
a=1, R^2=.94

Ham Radio

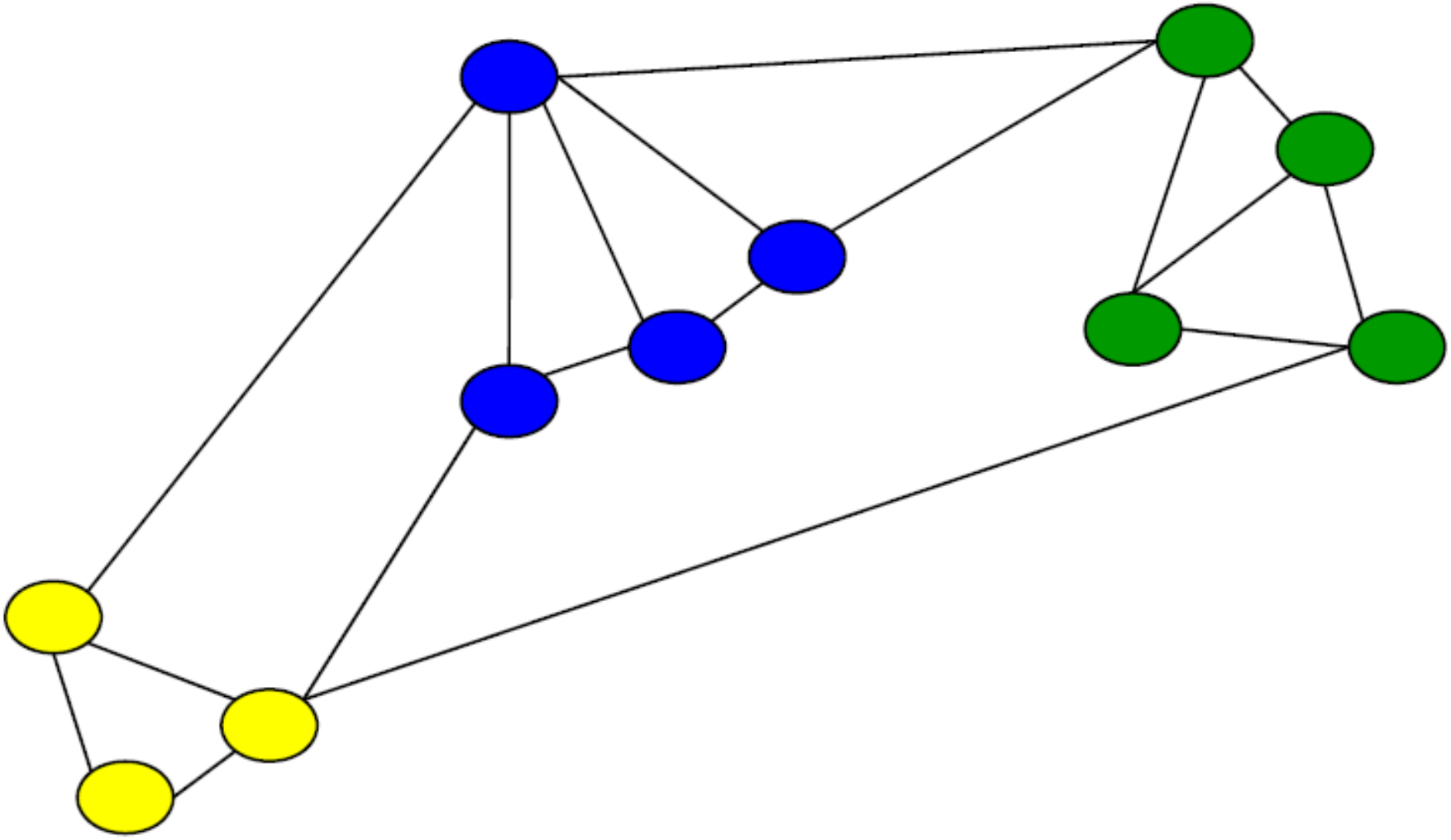


a=1, R^2=.94

High School Romance



a=1, R^2=.99



Блочные модели

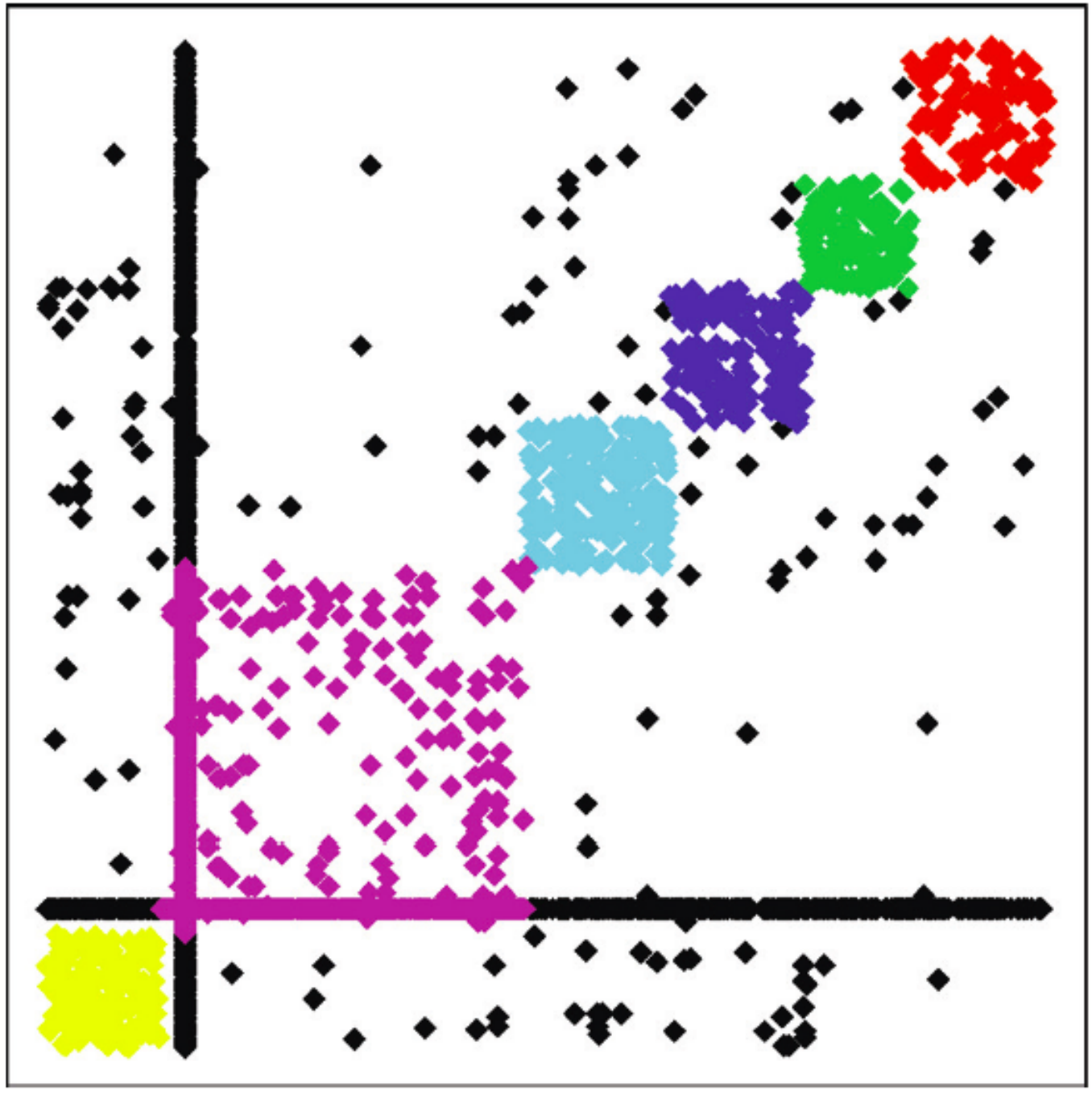
Блочные модели (Comola and Fafchamps, 2013) – обобщение модели Эрдеша-Реньи на случай, когда характеристики агентов влияют на вероятности.
Можно пытаться изучать связи вида:

$$\log \frac{p_{ij}}{1-p_{ij}} = \beta_i x_i + \beta_j x_j + \beta_{ij}|x_i - x_j| \tag{11}$$

В итоге это записывается в модель:

$$P(g|\boldsymbol{x}) = \prod_{i<j} p(x_i,x_j)^{A_{ij}}(1-p(x_i,x_j))^{1-A_{ij}} \tag{12}$$

Что, если мы не наблюдаем все характеристики и блоки или есть гетерогенность на индивидуальном уровне?



ERG модели

Пусть теперь вероятности образования ребер явным образом коррелированы и зависят от возможных структур на графе. Пример:

$$p = f(L, T) \tag{13}$$

Традиционно выражается как:

$$P_{\beta}(g) = \frac{\exp(\beta \cdot S(g))}{\sum_{g'} \exp(\beta \cdot S(g'))} \tag{14}$$

Чем хороши ERGM?

Теорема (Теорема Хаммерсли-Клиффорда)

Любая сеть может быть выражена через ERGM, если подобрать правильный набор статистик S.

- К S можно отнести произвольные статистики для графа и вершин
- ERGM дают фантастическую гибкость для подбора параметров

Как оценивают ERGM?

Есть два основных подхода:

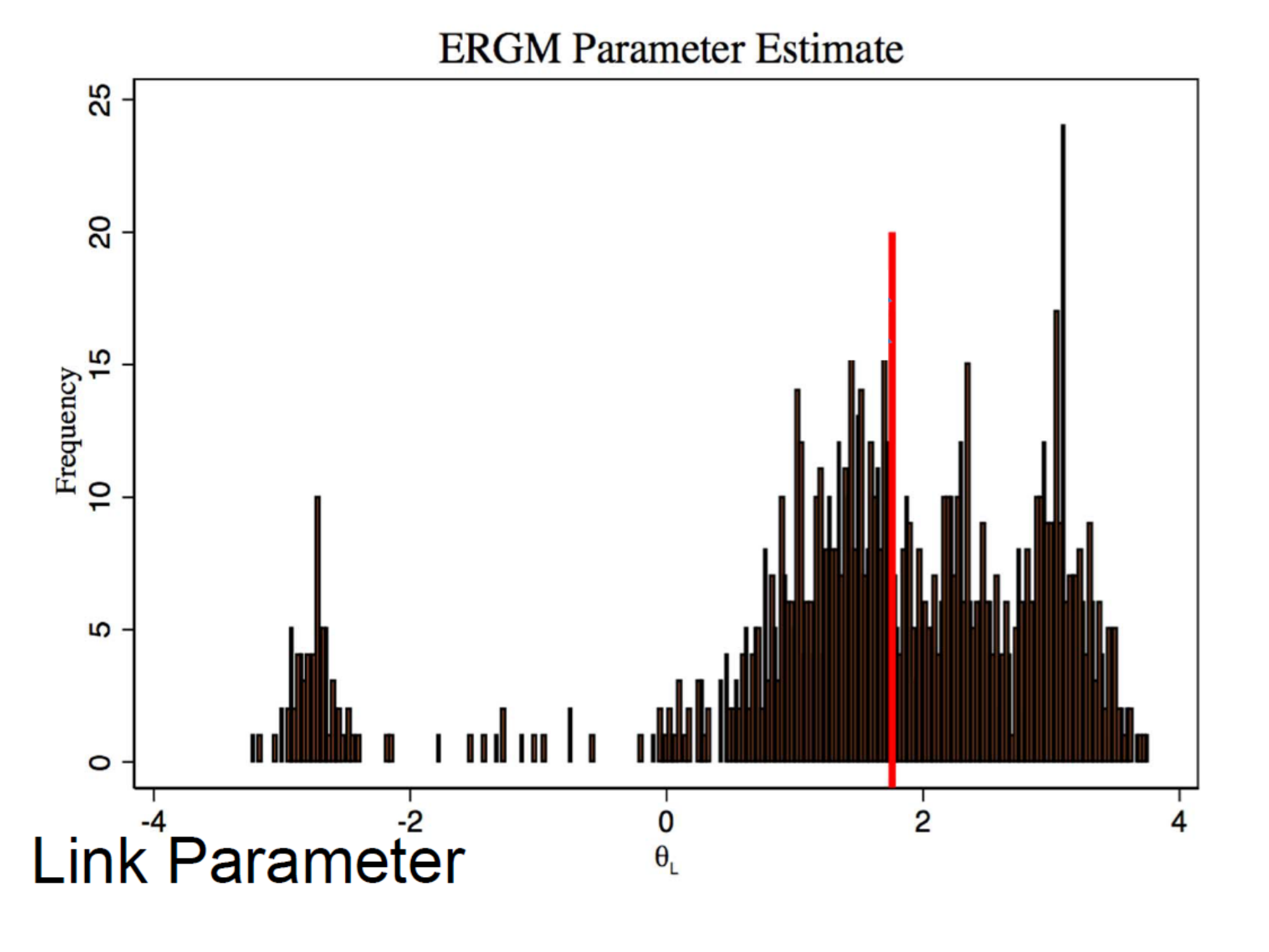
- В теории: метод максимального правдоподобия. На практике не используется, потому что знаменатель не может быть реалистично вычислен прямо
- Вариационные принципы (пример у Chatterjee, Diaconis (2013))
- MCMC: имплементация в пакете `statnet` в R

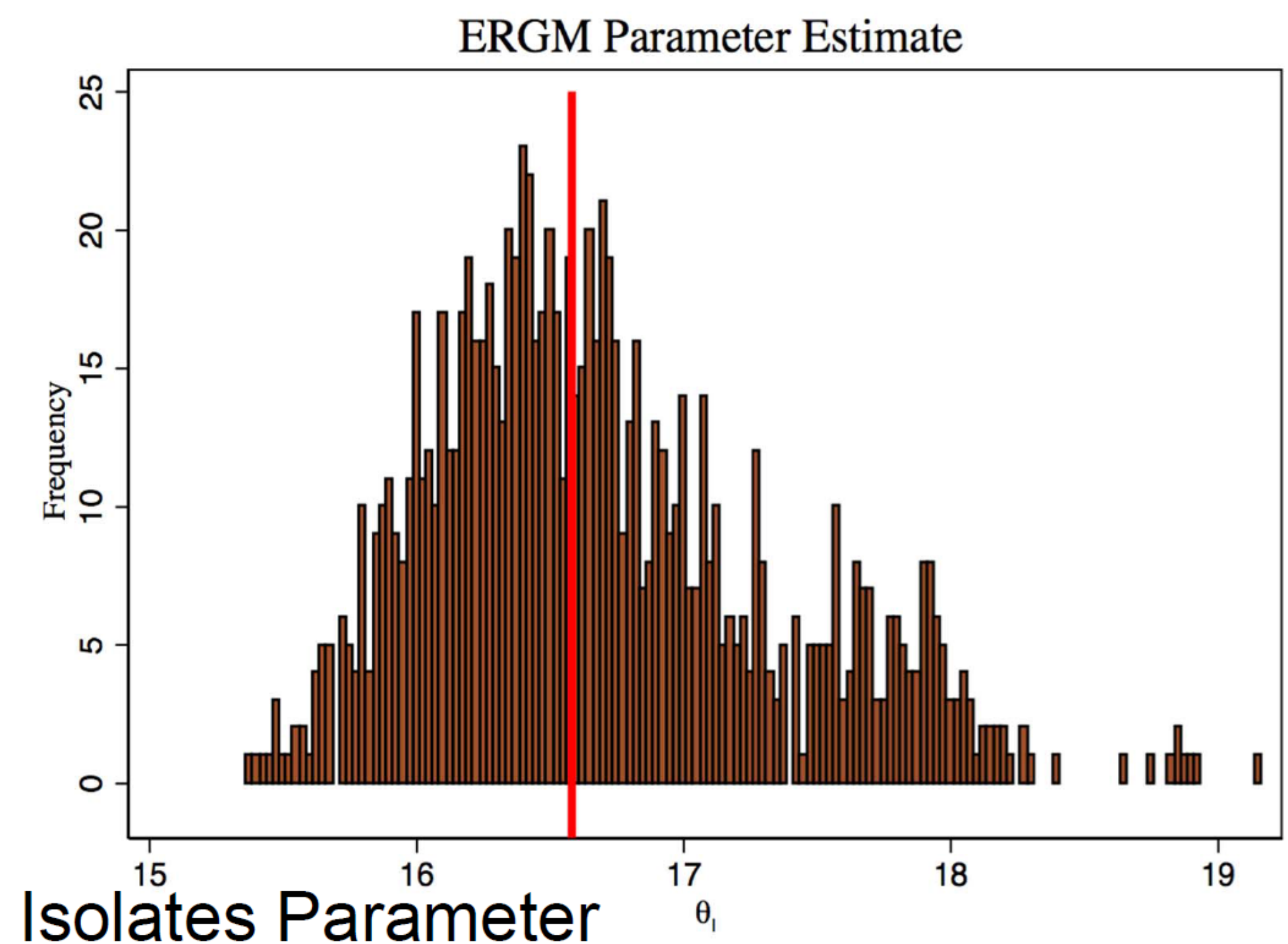
Проблемы с оценкой ERGM

Очень существенные проблемы в оценке таких моделей:

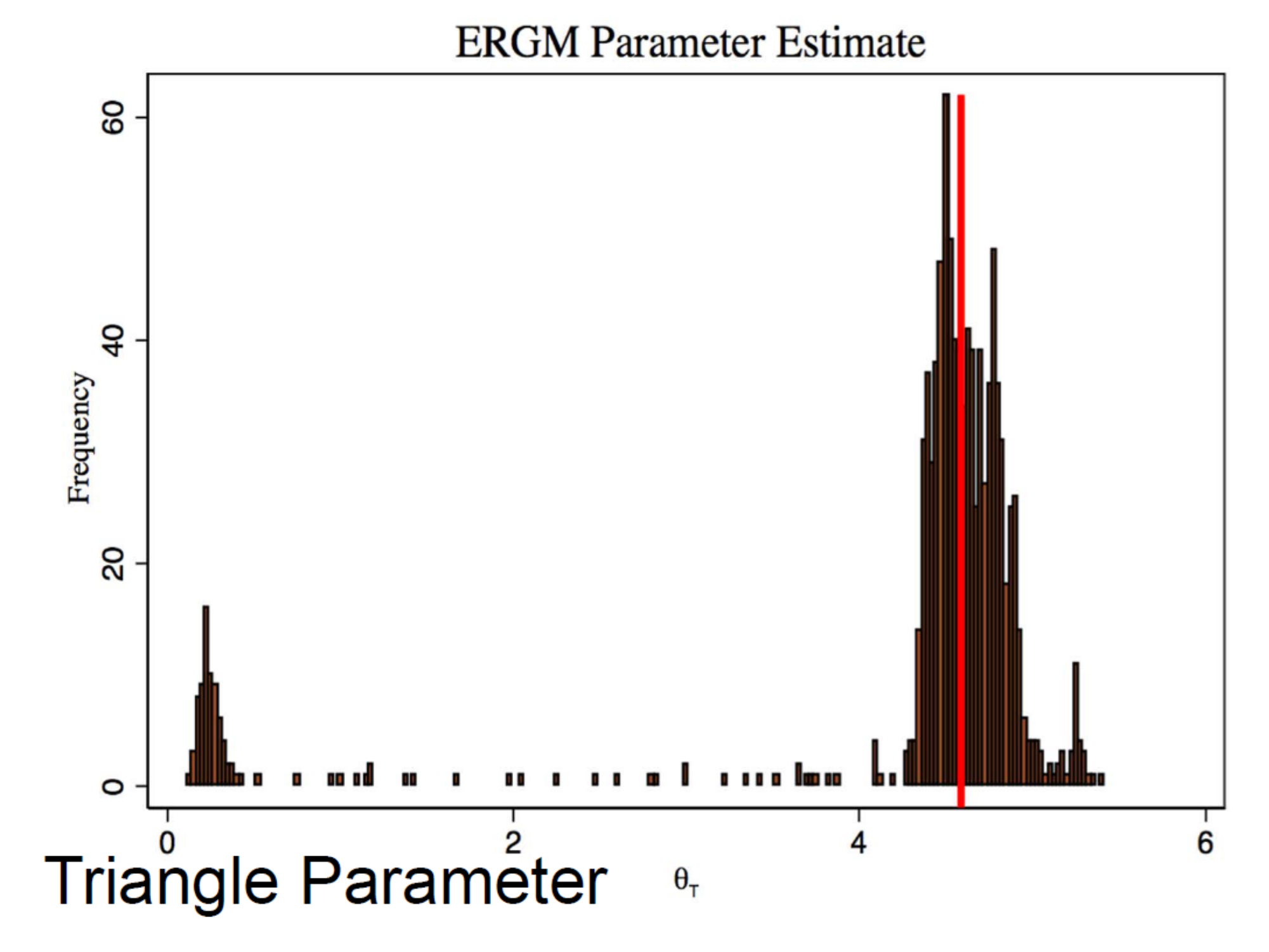
- Краевые вырожденности цепи (Snijders, 2002): бимодальные или полимодальные распределения, резкие скачки параметров
- Экспоненциальное время для сходимости во многих регионах, полиномиальное время только в регионах, где ребра почти независимы (что делает применение ERGM совершенно бессмысленным)

Пример симуляции: $n = 50$, $t = 1000$, 20 изолированных, 10 треугольников, 45 ребер


ERGM Parameter Estimate
Frequency
0
5
10
15
20
25
-4
-2
0
2
4
θL
Link Parameter



Isolates Parameter


ERGM Parameter Estimate
Frequency
0
20
40
60
0
2
4
6
Triangle Parameter
θ_T

SERGM

Чтобы сократить размерность, схлопываем все «одинаковые» графы. Получаются статистические ERG модели:

$$P_\beta = \frac{N(s)\exp(\beta \cdot S)}{\sum_{s'} N(s')\exp(\beta \cdot S)} \tag{15}$$

Суммирование идет уже по классам.
А хватит ли нам информации? Chandrasekhar и Jackson (2014) дают условия, при которых с ростом n оценки ММП сходятся к истинным значениям.

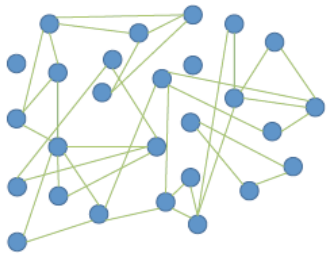
SUGM

Идет генерация по подграфам: треугольники, потом звезды, потом одиночные ребра. Интуиция: агенты постепенно организуют некие связи разных форм, образуя в итоге сеть.

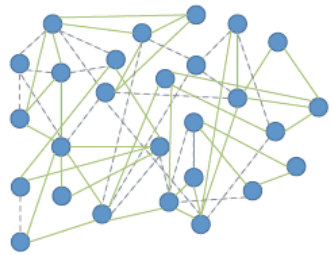
Главная проблема: наблюдаем только итоговую сеть. Если она плотная, почти невозможно оценить адекватно параметры. Большая часть наблюдаемых экономических сетей разреженная (Chandrasekhar (2015) дает оценки для этого).

Секвенциальная оценка параметров: выбрать структуру, найти все образцы, оценить логит-модель для нее, убрать все такие структуры. Повторять, пока не останется пустая сеть.

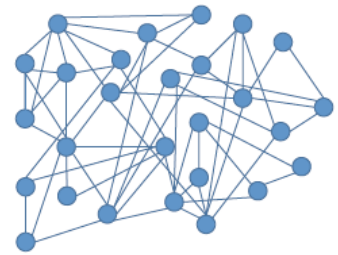
(A) n nodes



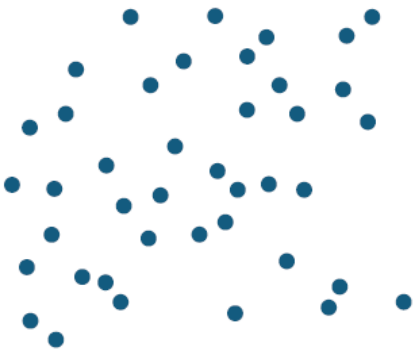
(B) Triangles form



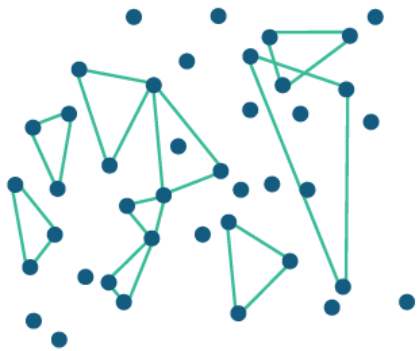
(C) Links form



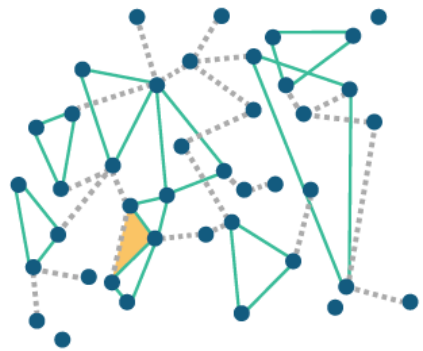
(D) Network



(E) n nodes



(F) Triangles form



(G) Links form

(H) Network

Эконометрическая версия connections game

Самая простая возможная игра стратегического формирования сети. Jackson, Wolinsky (1996):

$$U_i(g) = \sum_{j \in 1,\ldots,n, j \neq i} \delta^{d(i,j;g)-1}(1 + \varepsilon_{ij}) - |N_i(g)| \qquad (16)$$

Вопрос

Как оценить параметры?

Выбрать solution concept: обычно pairwise stability (Jackson, Wolinsky, 1996). Miyauchi (2016) работает с симуляциями, чтобы найти максимально «подходящие параметры».

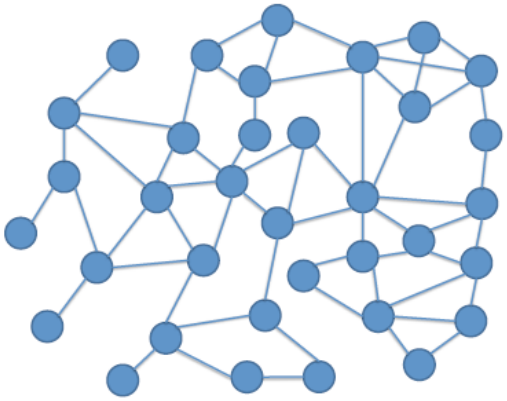
Проблема смещенности данных

В случае использования выборочных данных мы начинаем терять структуру и каналы влияния. Даже если оценки состоятельные, они становятся все более смещенными со снижением sampling rate.
Полное изучение проблемы – Chandrasekhar, Lewis (2015)
Рассматриваем регрессию исходов на характеристики сети:

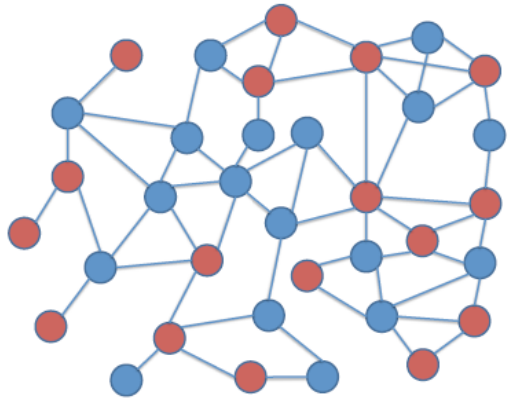
$$y_r = w_r \beta_0 + \varepsilon_r \tag{17}$$

Если есть ошибка измерения, то имеем регрессор $\tilde{w}$:

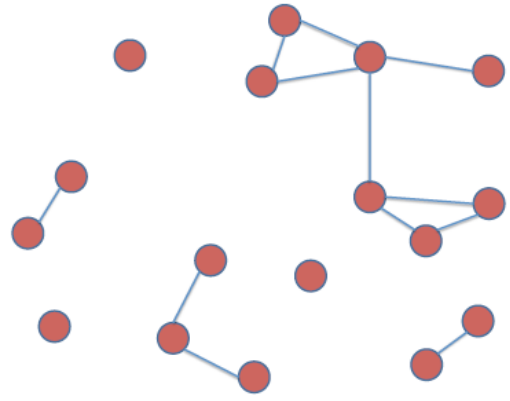
$$\text{plim}\hat{\beta} = \beta_0 \frac{\text{Cov}(\tilde{w}, w)}{\text{Var}(\tilde{w})} \tag{18}$$



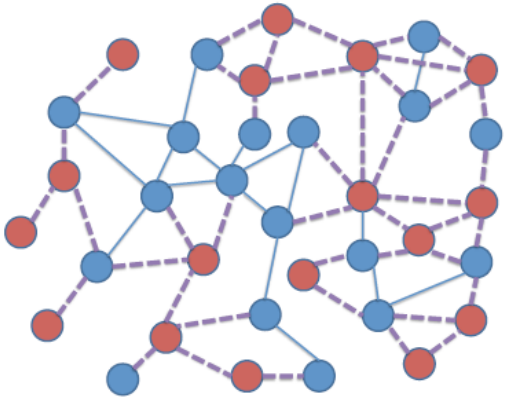
(A) Graph G



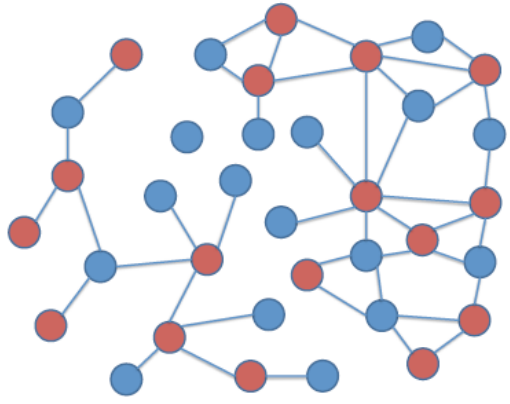
(B) Sampled nodes S



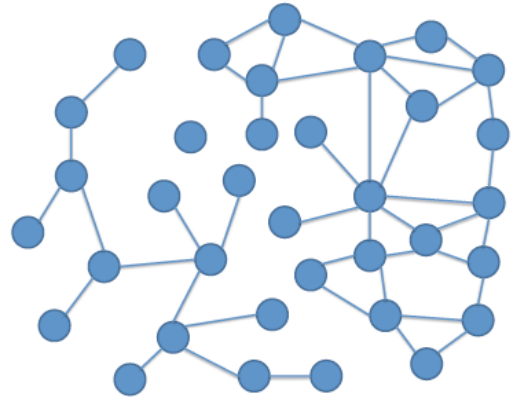
(C) Induced subgraph



(D) Highlights links used in star subgraph



(E) Star subgraph



(F) Star subgraph

Два способа коррекции

Chandrasekhar, Lewis (2015) предлагают два способа для преодоления проблемы:

- Аналитическая коррекция: для каждой характеристики сети надо подбирать отдельную коррекцию.
  Минус в том, что приходится строить новые сложные статистики для каждой статистики заново.
- Двухшаговая графическая реконструкция: универсальный способ.
  1. Оценить по данным распределение, из которого извлечены сети
  2. На основе этих теоретических данных оценить

Введение в эконометрику сетей

Владислав Морозов

27 декабря 2016 г.