Чем отличается настоящая карта контактов ДНК от случайной матрицы

Валерия Ковалева, Сергей Нечаев
valeriya.kovaleva@phystech.edu

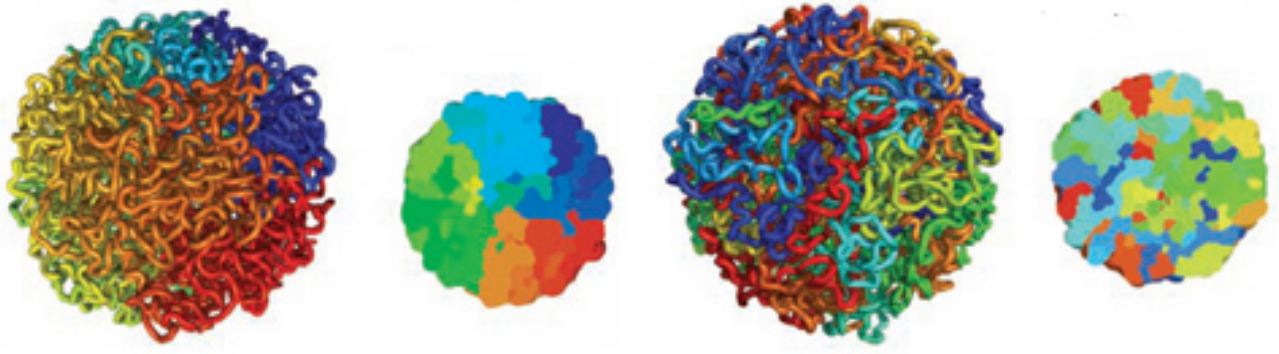
МФТИ, Сколтех

Май 2017

The unfolded DNA strand

folds into this
(FRACTAL GLOBULE)

rather than this
(EQUILIBRIUM GLOBULE)



Введение

Карта контактов - симметричная бинарная матрица с нулевой диагональю, где единица означает наличие контакта, а ноль - его отсутствие. Например, составленная так:

$$A_{ij} = \mathbb{1}\{r_{ij} < r_{\text{хар}}\},$$

где r_{ij} - расстояние между i и j мономером в пространстве, а $r_{\text{хар}}$ - характерное расстояние между мономерами.

Можно рассматривать индивидуальную карту, сумму нескольких и сумму большого числа таких карт. Все такие карты будем называть картами контактов.

Введение

У карты контактов будем смотреть:

- спектральную плотность,
- устойчивость иерархии при нахождении доменов с помощью кластеризации взвешенного графа,

и сравнивать со случайными матрицами.

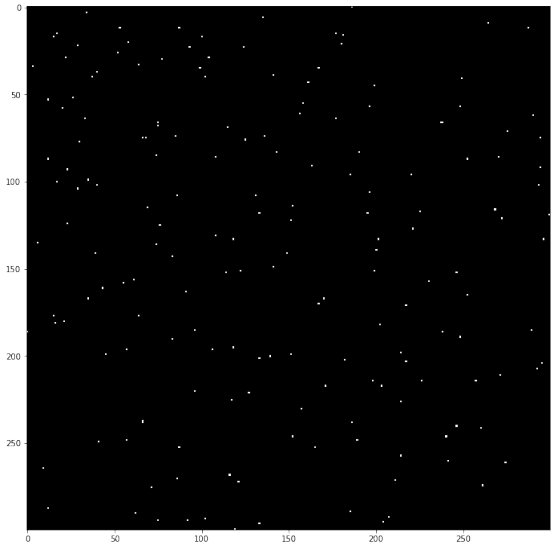
Отдельно можно смотреть спектральную плотность лапласовской матрицы:

$$\mathbf{L} = \mathbf{D} - \mathbf{A},$$

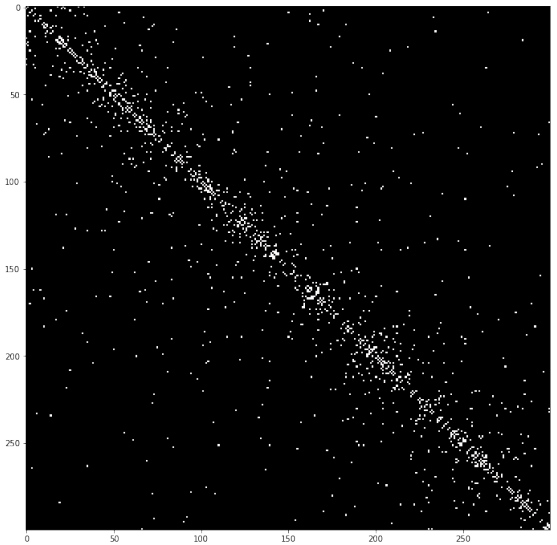
где $\mathbf{D}$ - диагональная матрица валентностей, $\mathbf{A}$ - карта контактов.
Собственные значения лапласовской матрицы имеют топологический смысл.

Случайные матрицы

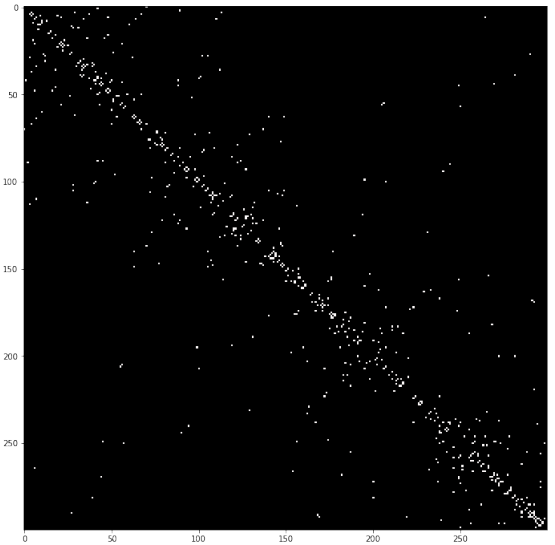
- матрица смежности случайного графа Эрдеша-Реньи,
- матрица с элементами $X_{ij} \sim \frac{c_\alpha}{|i-j|^\alpha}$,
- матрица из блоков, сгенерированных одним из двух предыдущих способов с различными параметрами.



Эрдеш-Реньи



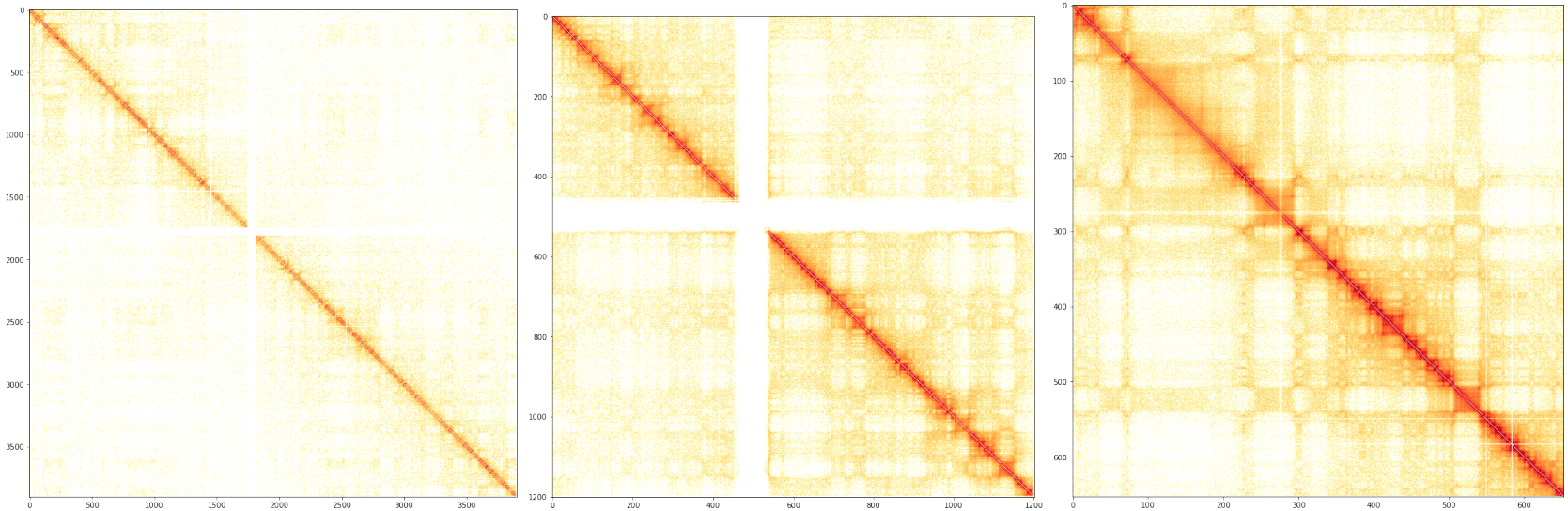
$X_{ij} \sim \frac{c_\alpha}{|i-j|^\alpha}$



Блочная

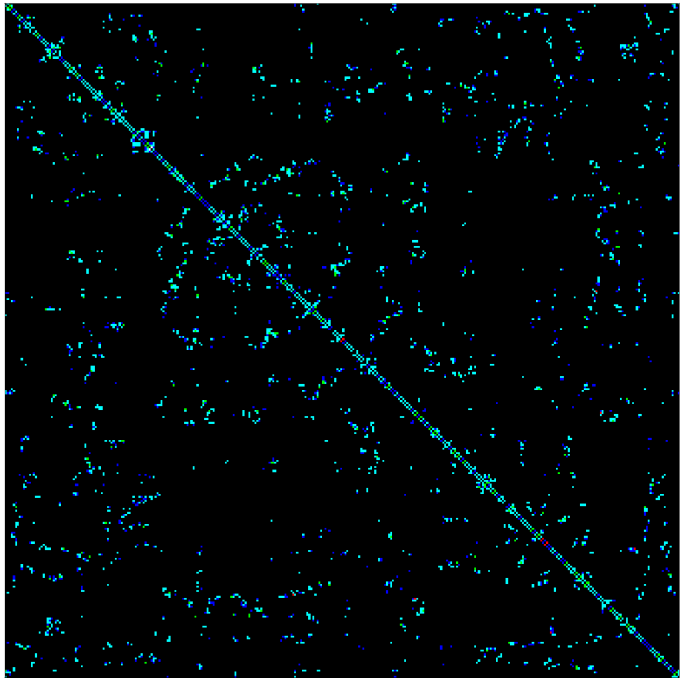
Hi-C карты

- в разрешении 1 кб, но мы работали в основном с 50 кб,
- разреженные и “плотные” (по большому числу измерений),
- имеют ярковыраженную “шахматную” структуру,
- разного происхождения: человека, мыши и без указания.

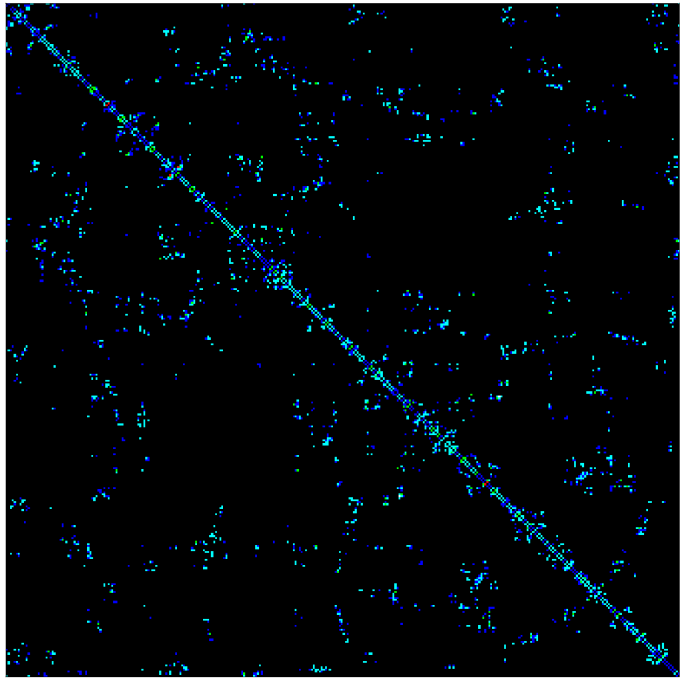


Полусинтетические глобулы

Данные Максима Имакаева по коллапсу длинных цепочек в ящике с достаточно большой плотностью:



equilibrium



mixed

Кластеризация

Идея

Карте соответствует граф, который можно разбить на кластеры. Мономеры внутри одного домена больше взаимодействуют между собой и соответствуют рёбрам с большими весами, а значит, будут определены в один кластер.

Метод

Решаем задачу дискретной оптимизации - максимизации функционала модулярности:

$$Q[\mathbf{W}, \mathbf{C}] = \frac{1}{2w} \sum_{i,j} \left(w_{ij} - \frac{w_i w_j}{2w} \right) \sum_{\alpha} c_{i\alpha} c_{j\alpha} \to \max_{\mathbf{C}}$$

Функционал модулярности

Идея

Перебирает всевозможные разбиения по кластерам, сравнивает, насколько больше связей в таком подграфе, чем было бы в подграфе графа Эрдеша-Реньи.

Краткая характеристика

- хорошо подходит для нашей задачи,
- в контексте модели можно решать жадным поиском вместо полного перебора,
- имеет предел разрешения $\sim \sqrt{w}$, не находит кластеры с суммой рёбер меньше $\sqrt{w/2}$.

Модификация для уменьшения предела разрешения

$$\widetilde{\mathbf{W}} = \mathbf{W} + r\mathbf{I}$$

Алгоритм

Идея

Изменяя r в большом диапазоне, получим доменную иерархию - системы кластеров различных масштабов - и исследуем её устойчивость.

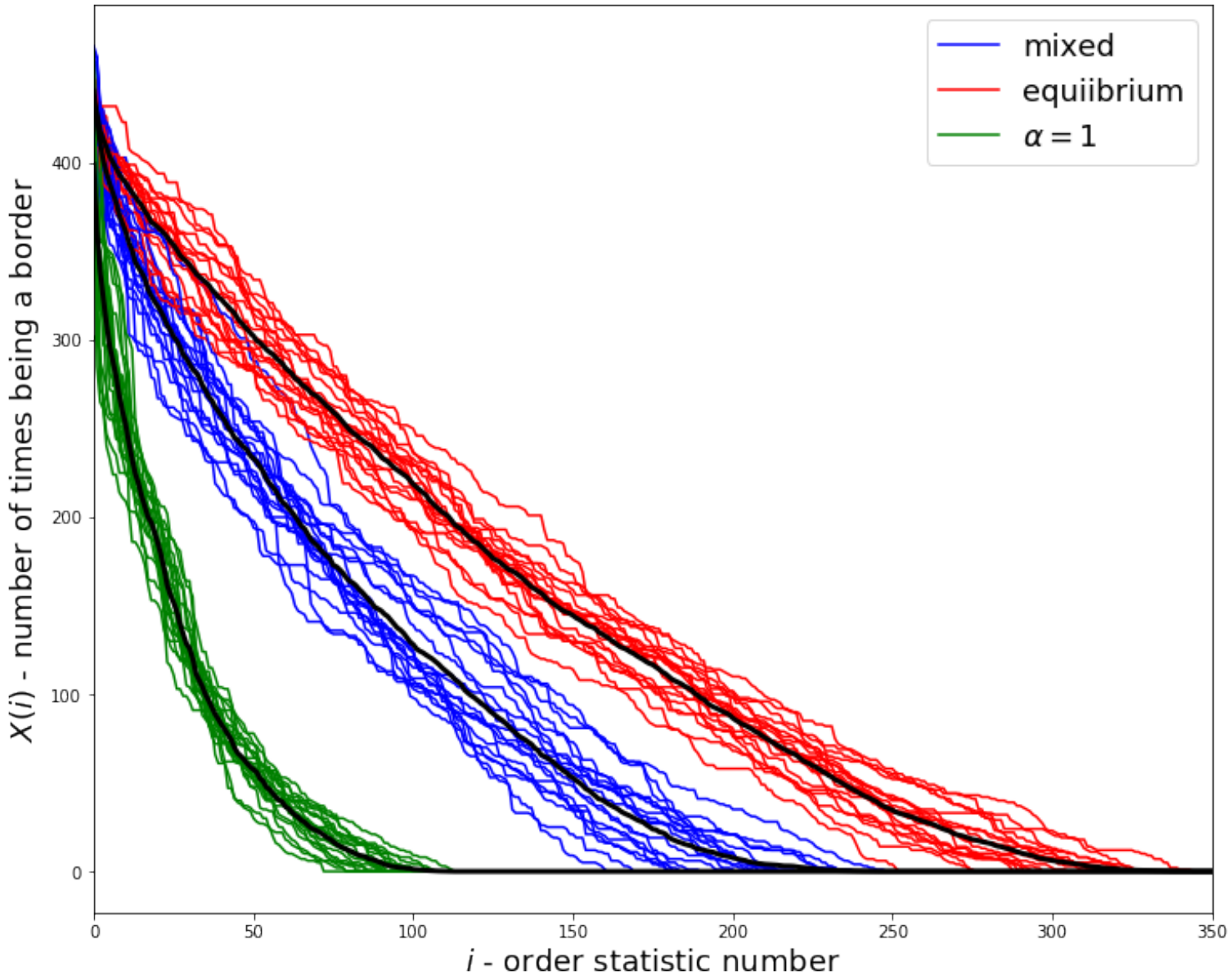
Предложение (Тамм, Москалец)

Построить для каждой вершины статистику - число раз, когда она являлась границей между кластерами. Чем устойчивее структура, тем полученная кривая будет ниже и более выпуклой.

Данные

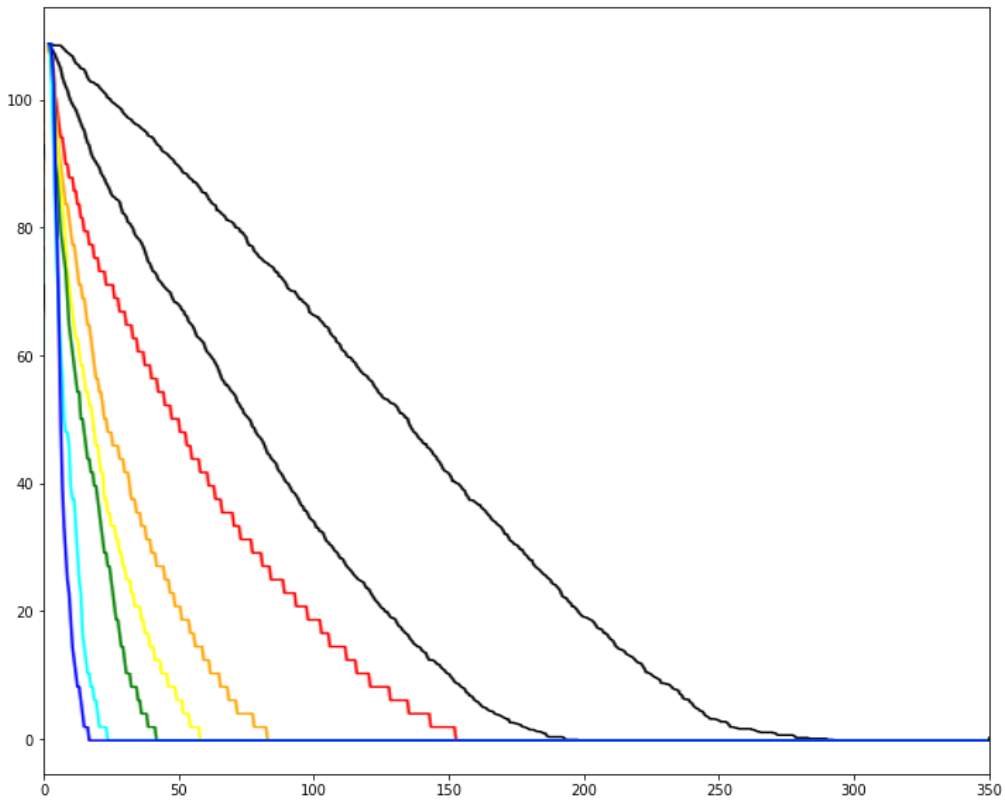
- глобулы,
- случайные матрицы с примерно тем же числом рёбер,
- Hi-C карты.

Результаты


mixed
equiibrium
$\alpha = 1$
$X(i)$ - number of times being a border
i - order statistic number
0
100
200
300
400
0
50
100
150
200
250
300
350

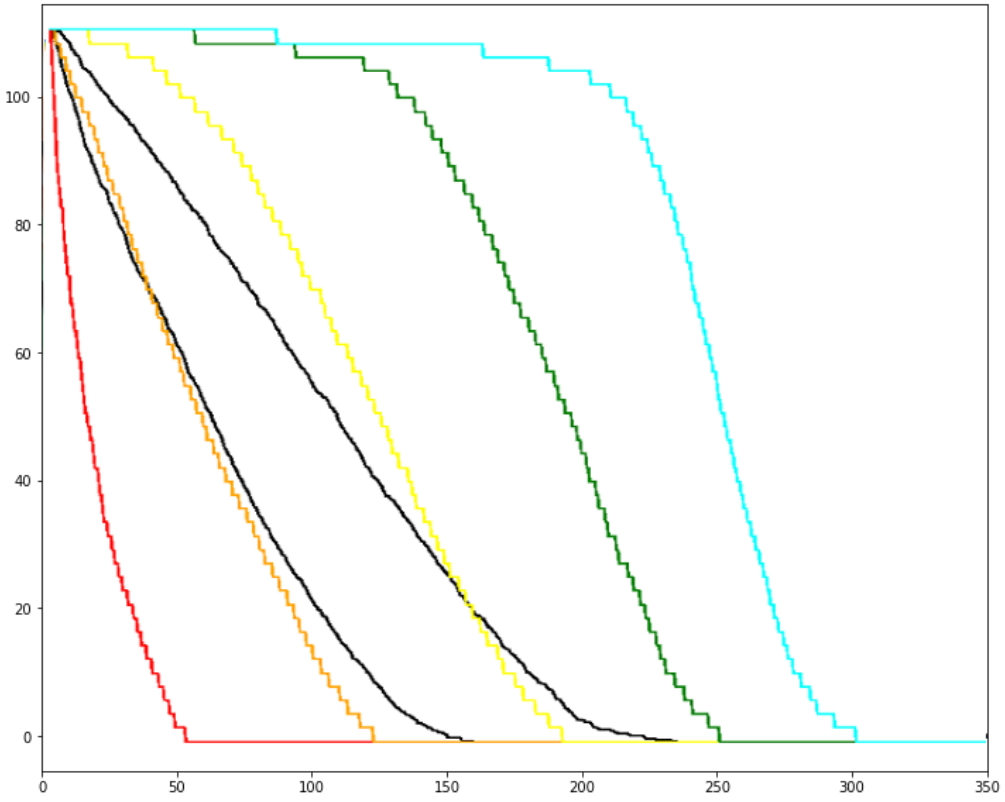
Сравнение кривых с фиксированными α и разными c

С увеличением c увеличивается и общее число связей, тем не менее, разбиение происходит более “неохотно”.
Здесь $\alpha = 1$, $c = 0.3,\ 0.5,\ 0.7,\ 1.0,\ 3.0,\ 5.0$.



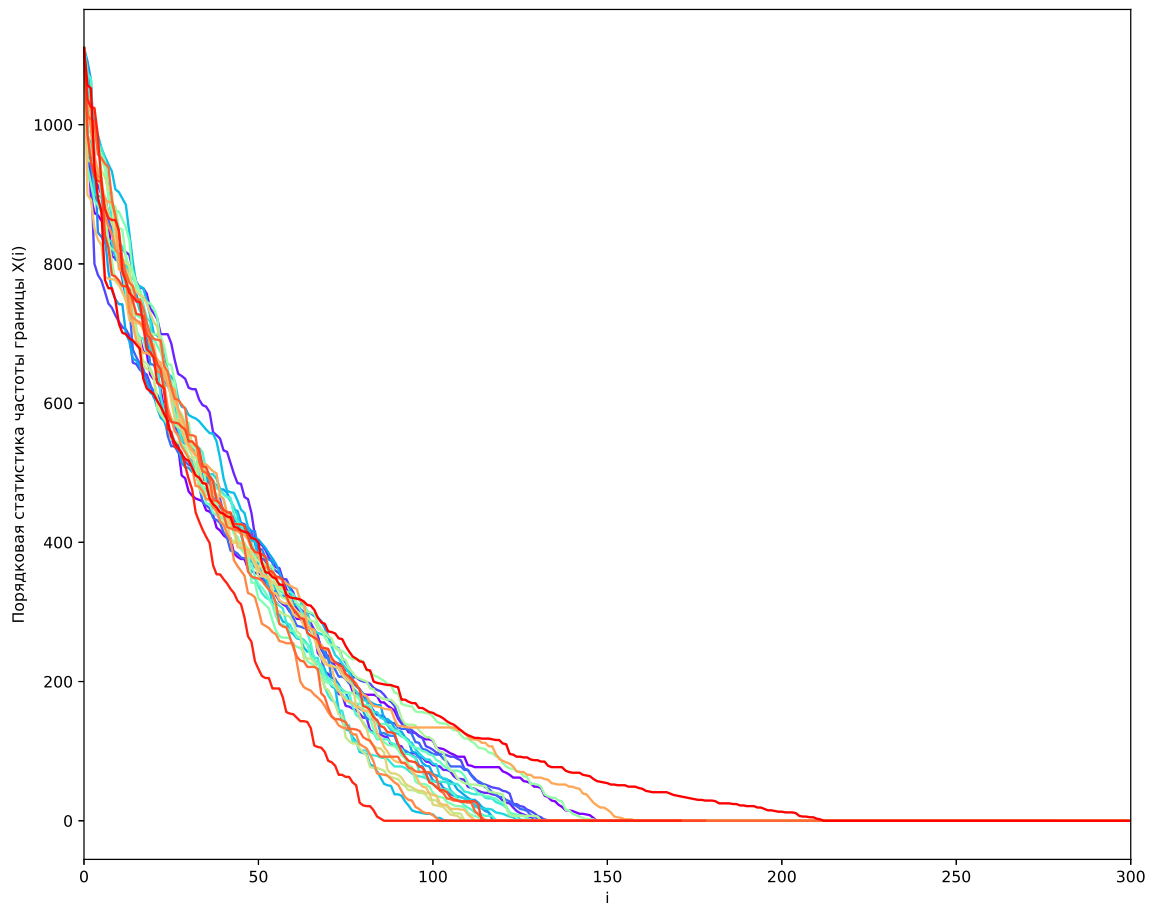
Сравнение кривых с фиксированным c и различными α

С увеличением α граф становится более разреженным, характер кривой катастрофически меняется.
Здесь $c = 0.66, \alpha = 1.0,\ 1.5,\ 2.0,\ 2.5,\ 3.0$



Hi-C

Легко разбиваются на кластеры, зависимость быстро спадает, ведут себя очень похоже **независимо от размера**.



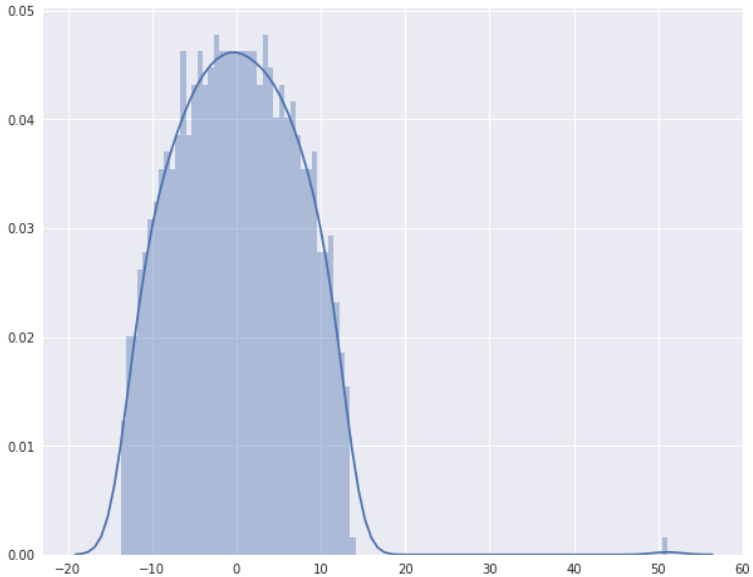
Спектральные плотности и распределение степеней

Спектральная плотность - распределение собственных чисел матрицы смежности.

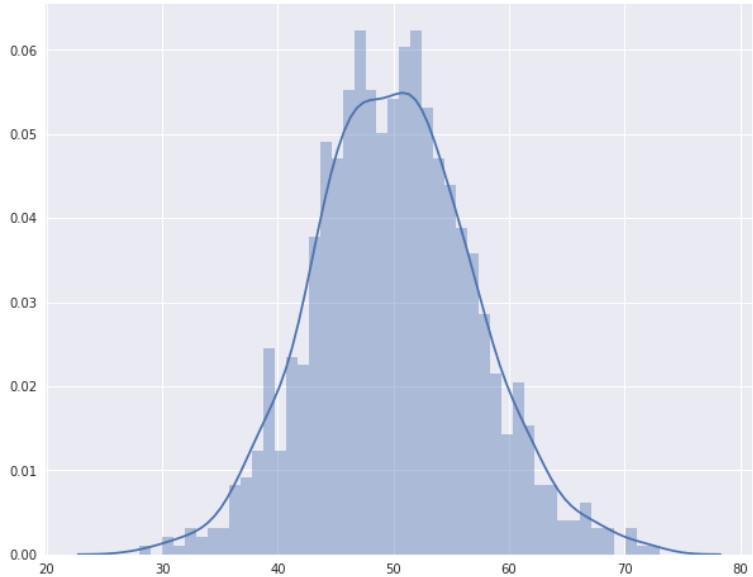
Спектральные плотности карт существенно различны у **всех** рассмотренных типов карт, но внутри одного типа очень похожи.

Случайные матрицы берутся с такими p и c, чтобы число связей в них было примерно равно настоящим картам.

Эрдеш-Реньи

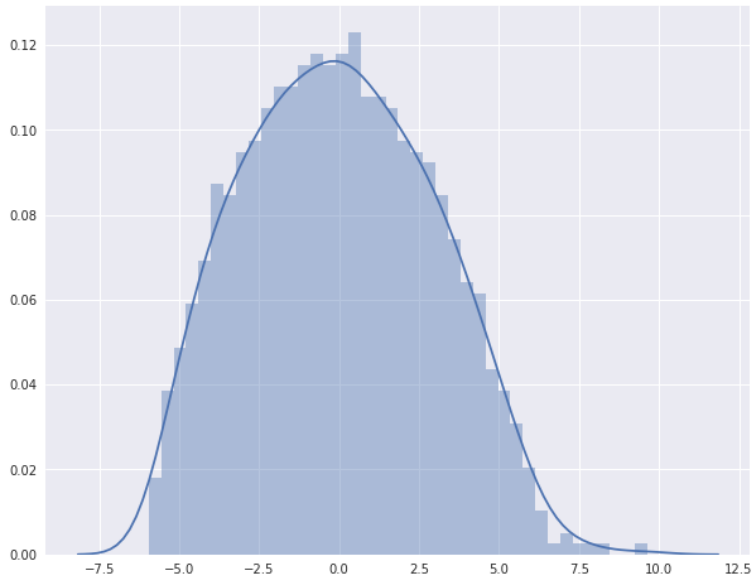


Спектр имеет форму полукруга + убегающее значение

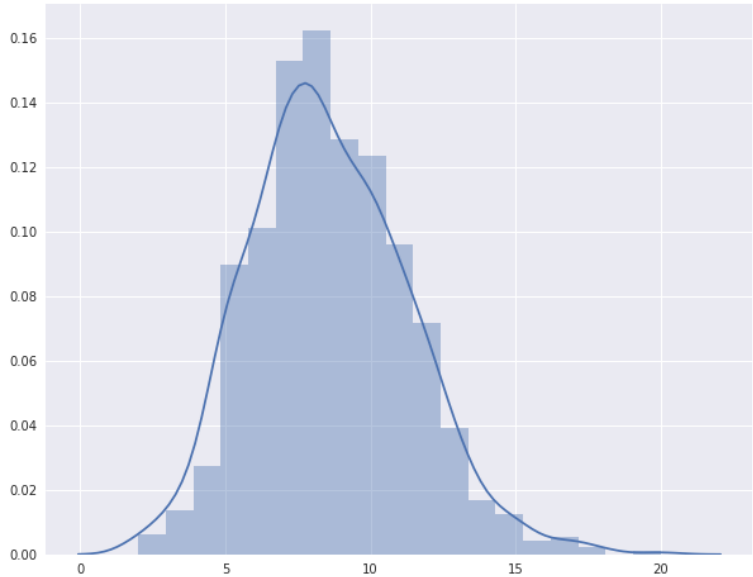


Валентности распределены по Пуассону

Матрицы с диагональным убыванием

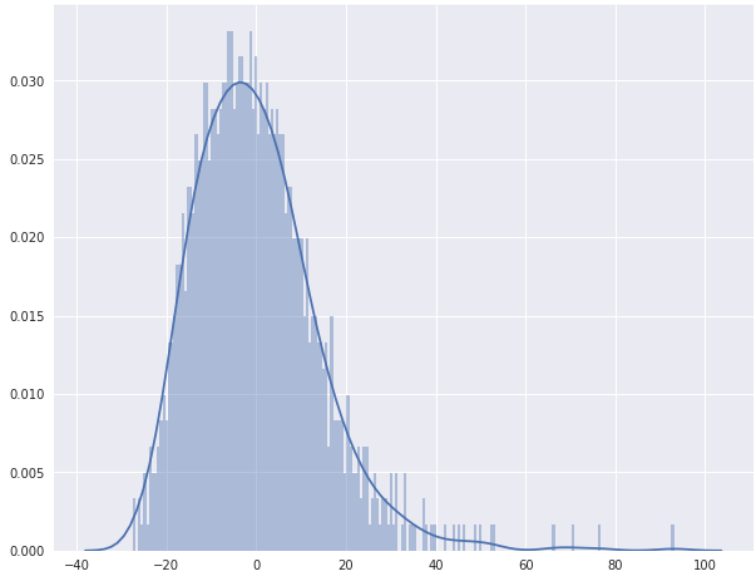


“Левая” половина спектра - полукруг, “правая” имеет хвост

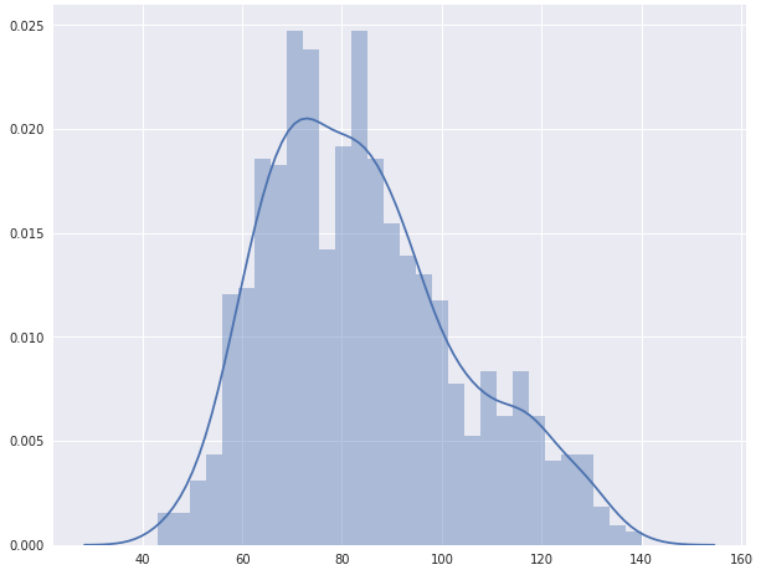


Распределение валентностей всё ещё похоже на Пуассона

Матрицы из блоков



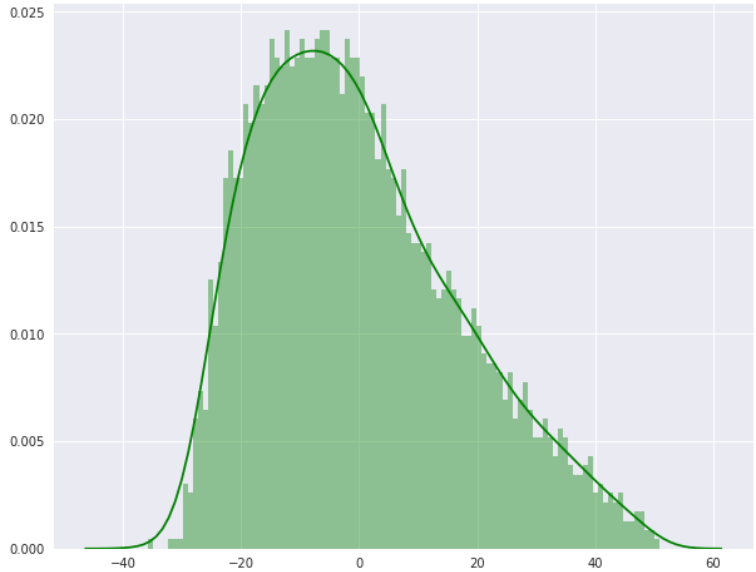
У спектра появляется хвост слева, утяжеляется хвост справа



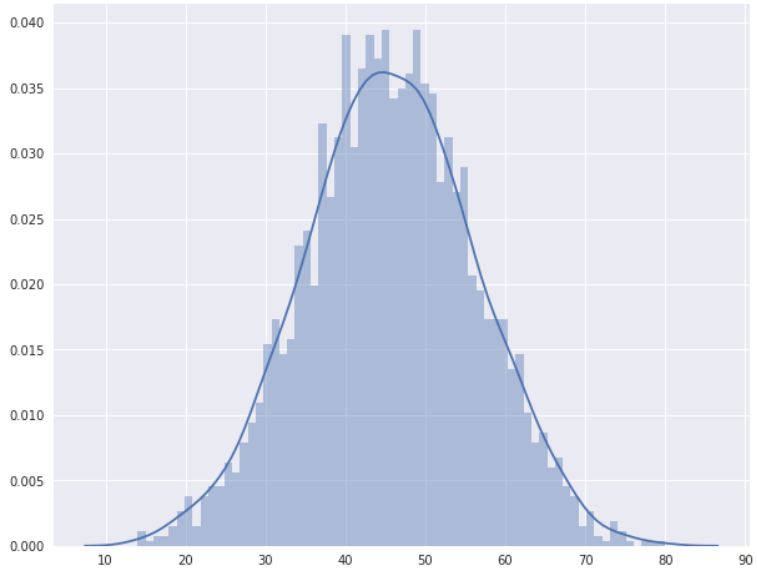
Распределение валентностей имеет загадочный вид

Глобулы

Спектральные плотности и распределения степеней вершин выглядят почти одинаково для искусственных equilibrium и mixed глобул.

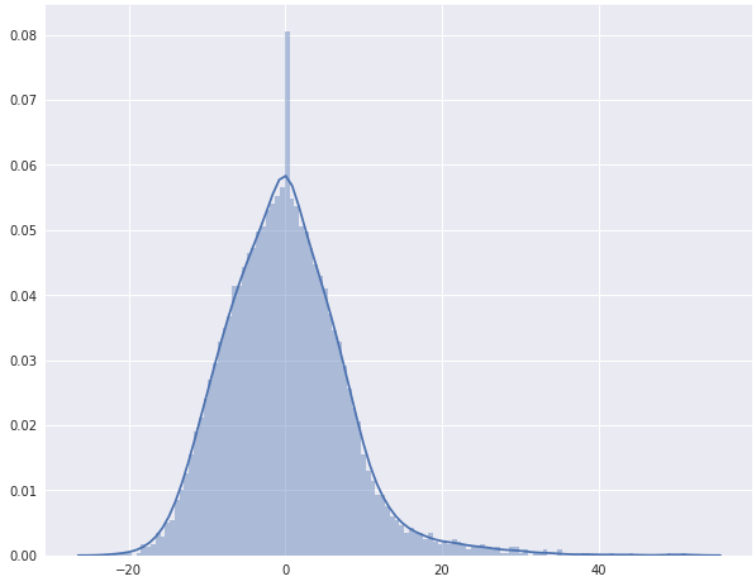


Спектральная плотность отличается от всех предыдущих

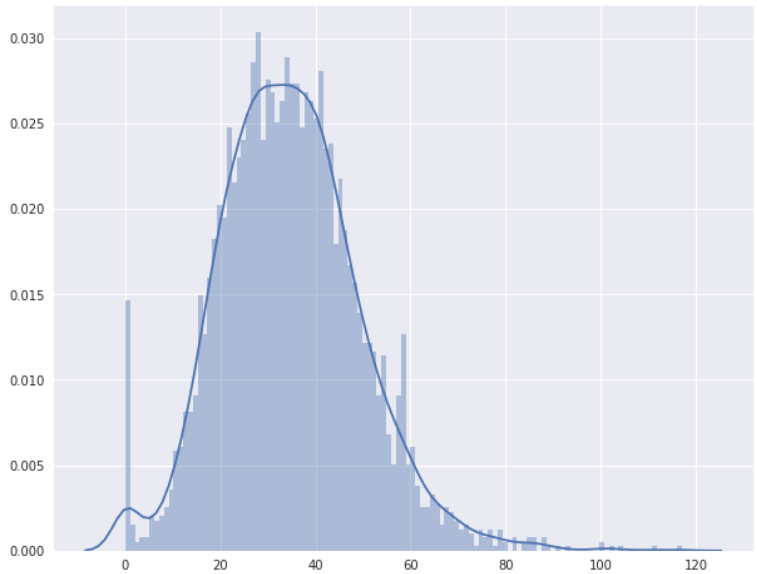


Распределение валентностей пуассоновсское

Разреженные Hi-C

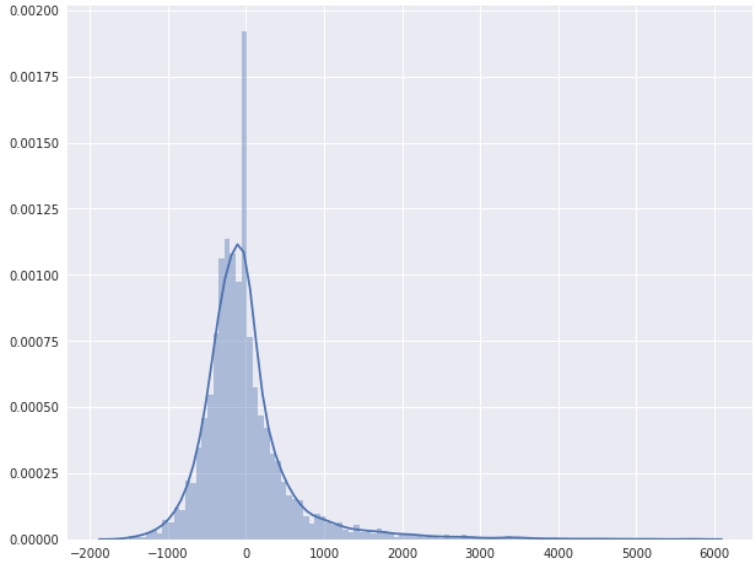


Спектральная плотность треугольная с тяжёлыми хвостами и выбросом в нуле

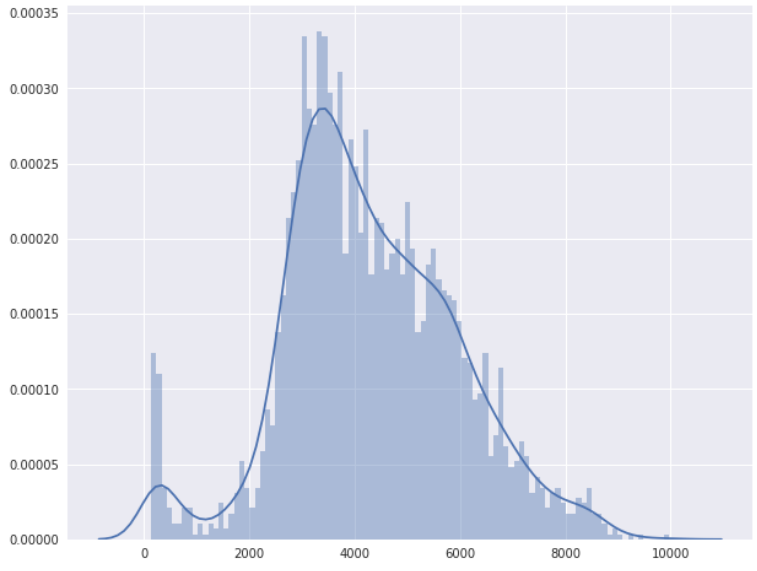


Распределение вершин всё ещё пуассоновское

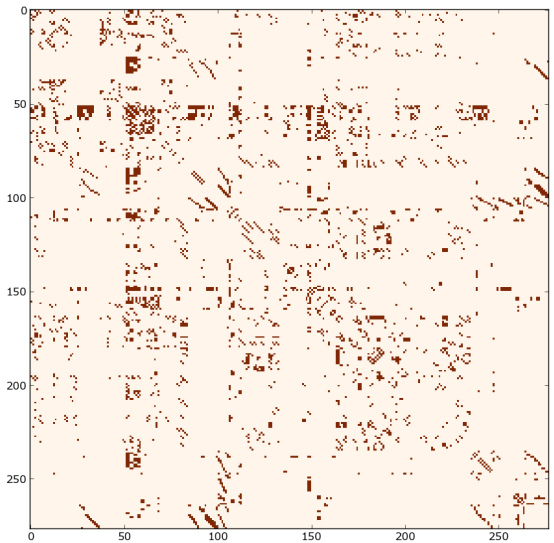
Обычные Hi-C



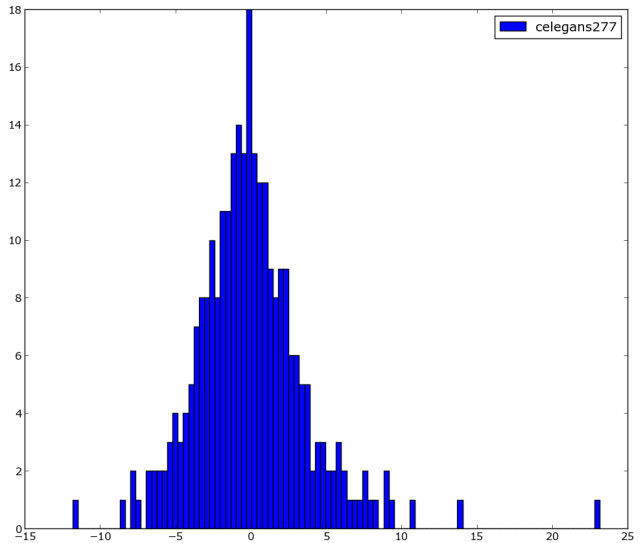
У спектральной плотности также тяжёлые хвосты и выброс в нуле



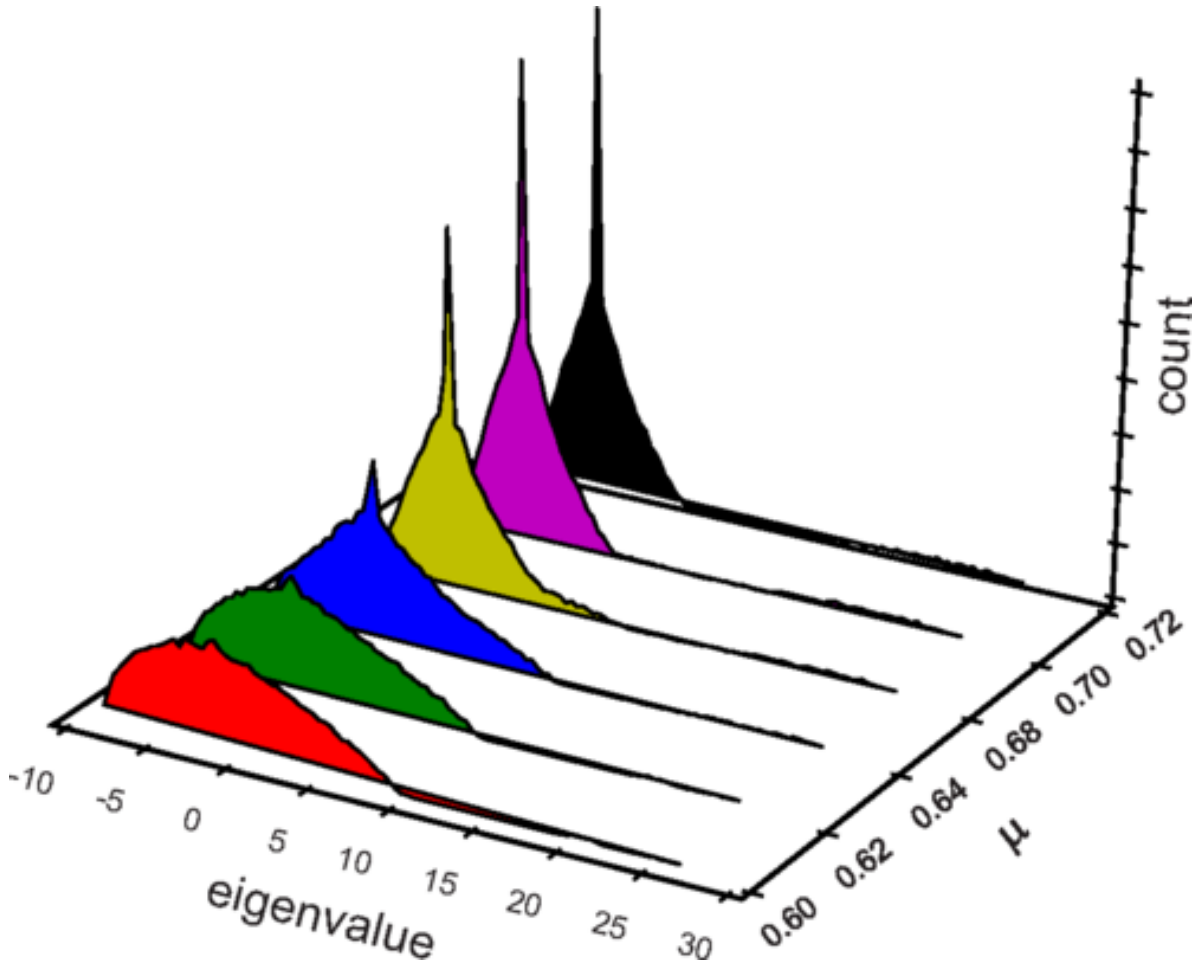
В распределении валентностей появляются “ступеньки”



Карта синаптических связей похожа на Hi-C карту



Спектральная плотность похожа на плотность для Hi-C



Эволюция спектра при сохранении валентностей

Заключение

- устойчивость границ существенно зависит от диагонального спадания,
- в настоящих картах границы более устойчивы, чем в случайных, а зависимость более выпуклая,
- поведение алгоритма похоже на Hi-C картах разного размера.

- важно смотреть на хвосты спектральных плотностей,
- правый хвост появляется из-за диагонального спадания, а левый - из-за “шахматной” структуры,
- также хвосты связаны с числом треугольников в соответсвующем графе,
- при этом распределение валентностей единообразно.