

Совместная встречаемость в русской и английской Википедиях (русские классические композиторы)

Технические данные.

NB! Все файлы находятся в репозитории проекта.

1. Репозиторий – <https://github.com/Smyek/DH-project>
2. Список русских классических композиторов был взят с этого сайта <http://www.greatcomposers.ru/index.php?go=Content&id=119> и хранится в файле **composers_ru.txt**
3. Файл с именами композиторов и их английским соответствием в **composers.csv** (создается при помощи нахождения ссылки на английскую версию на странице русской, т.к. модуль wikipedia 1.4.0 для python не поддерживает переключение языка страницы)
4. Файлы **%язык%_connections.csv** – файлы, содержащие массивы связей типа (source, target)

В качестве инструмента визуализации использовалась программа Gephi, файлы для которой создавались при помощи пакета networkx (python) на основе файлов из 4 пункта.

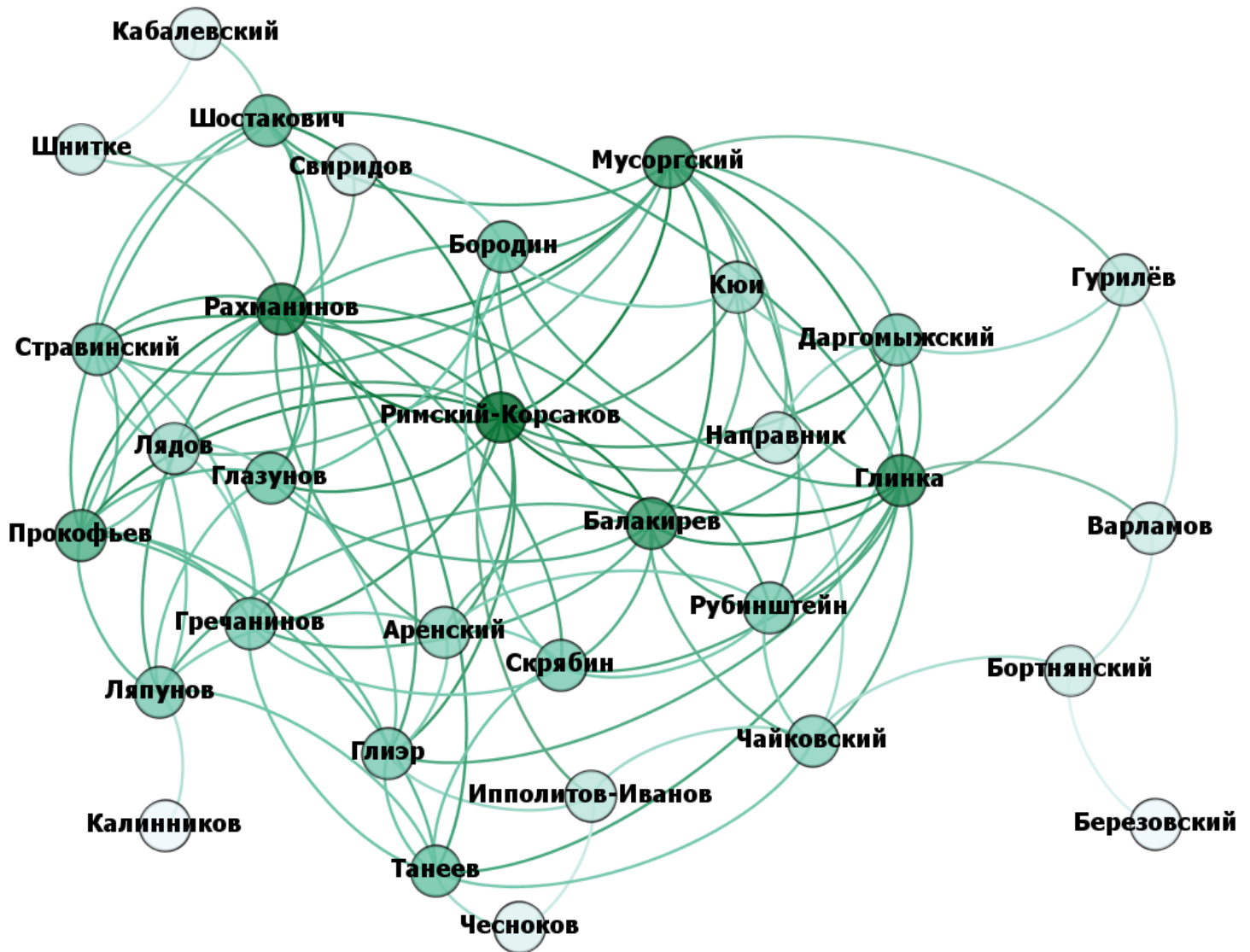
Degree – количество связей узла, можно считать рейтингом.

Modularity – метрика, показывающая как хорошо социальная сеть разбивается на сообщества узлов. Здесь нам скорее важно, то насколько они отличаются, а не сама оценка.

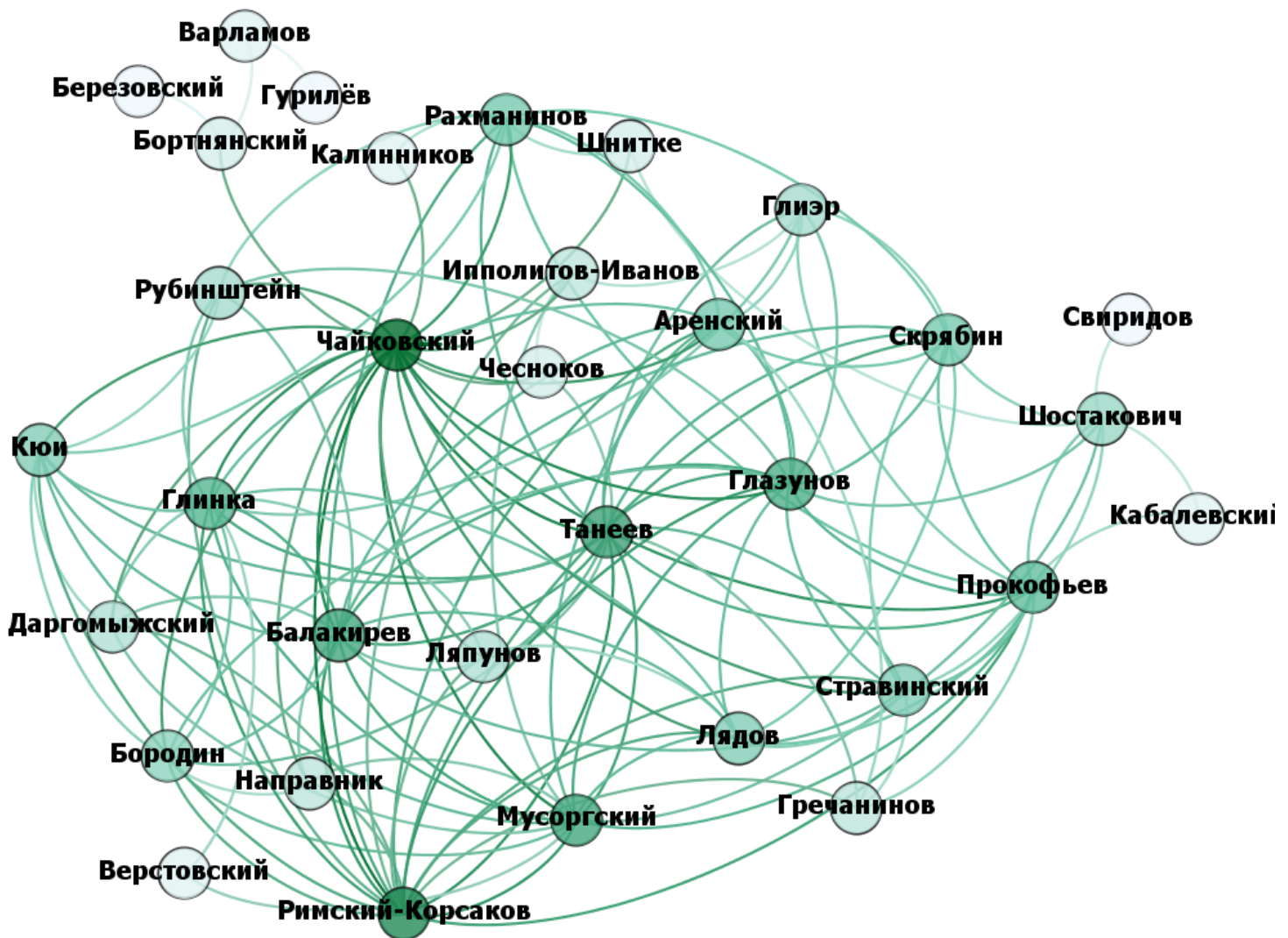
Визуализация и комментарии.

Посмотрим на социальные сети, получившиеся в итоге обработки страниц 34 русских классических композиторов.

Русская Википедия



Английская Википедия



Степень насыщенности цвета зависит от количества связей у ячейки. Мы видим множество сходств в связях, а также совпадения по degree узлов. Наиболее «популярные» Римский-Корсаков, Балакирев, Прокофьев и менее Безезовский, Бортнянский, Гурилев ведут себя примерно одинаково.

Больше всего бросается в глаза Чайковский: в английской Википедии он куда больше упоминается в статьях. Это также видно в табличке (ниже) сортированной по degree русских композиторов.

- В колонке diff считается разность degree узлов из ru и en Википедий.
- В diff_neighbour я попытался прикинуть функцию, которая считает такой дифф в зависимости от соседей, нормализуя и немного разглаживая различия, чтобы видеть где происходят несостыковки. Считается как $\log(\text{разность диффов соседних узлов} / \text{дифф самого узла})$.

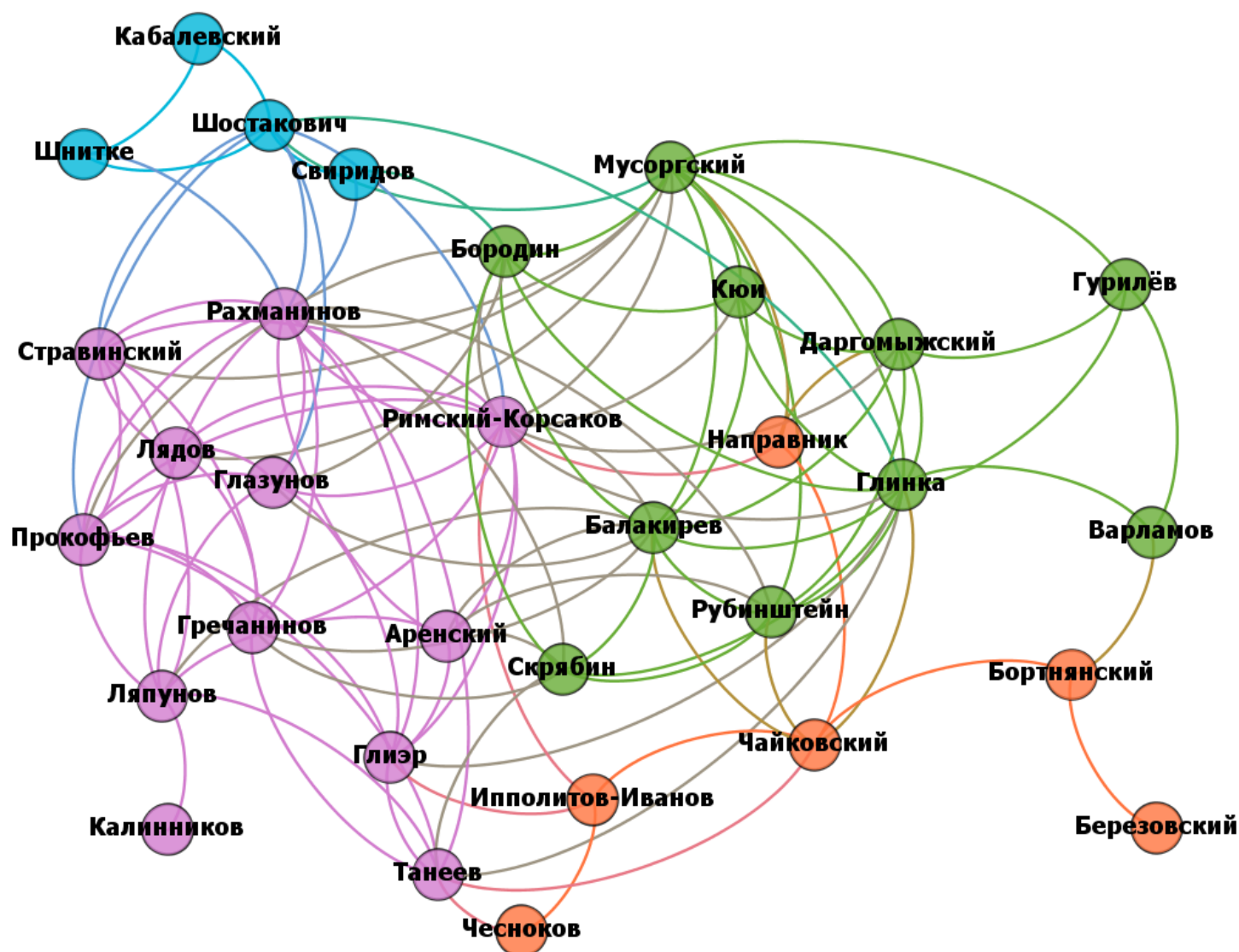
label/degree	ru	en	diff	diff_neighbour
Римский-Корсаков	17	20	3	0,8
Рахманинов	15	11	4	0,8
Глинка	14	14	0	0,8
Балакирев	13	16	3	0,8
Мусоргский	13	16	3	0,9
Прокофьев	11	13	2	0,8
Шостакович	10	8	2	0,6
Бородин	9	9	0	0,9
Глазунов	9	15	6	1,1
Танеев	9	17	8	1,2
Глиэр	8	7	1	1,1
Гречанинов	8	5	3	0,8
Даргомыжский	8	6	2	0,8
Ляпунов	8	6	2	0,7
Рубинштейн	8	7	1	0,8
Скрябин	8	11	3	0,8
Стравинский	8	10	2	1,0
Аренский	7	11	4	1,4
Чайковский	7	24	17	1,4
Кюи	6	10	4	1,4
Лядов	6	10	4	1,0
Гурилёв	4	1	3	0,9
Ипполитов-Иванов	4	5	1	0,7
Направник	4	5	1	0,3
Бортнянский	3	3	0	0,3
Варламов	3	2	1	0,5
Свиридов	3	1	2	0,5
Шнитке	3	3	0	0,3
Кабалевский	2	2	0	0,0
Чесноков	2	3	1	0,0
Березовский	1	1	0	0,3
Калинников	1	2	1	0,0

Здесь же видно насколько похож низ рейтинга. Заедем, что Глазунов и Танеев также имеют более высокий degree в английской википедии, а последний вошел топ-5 рейтинга:

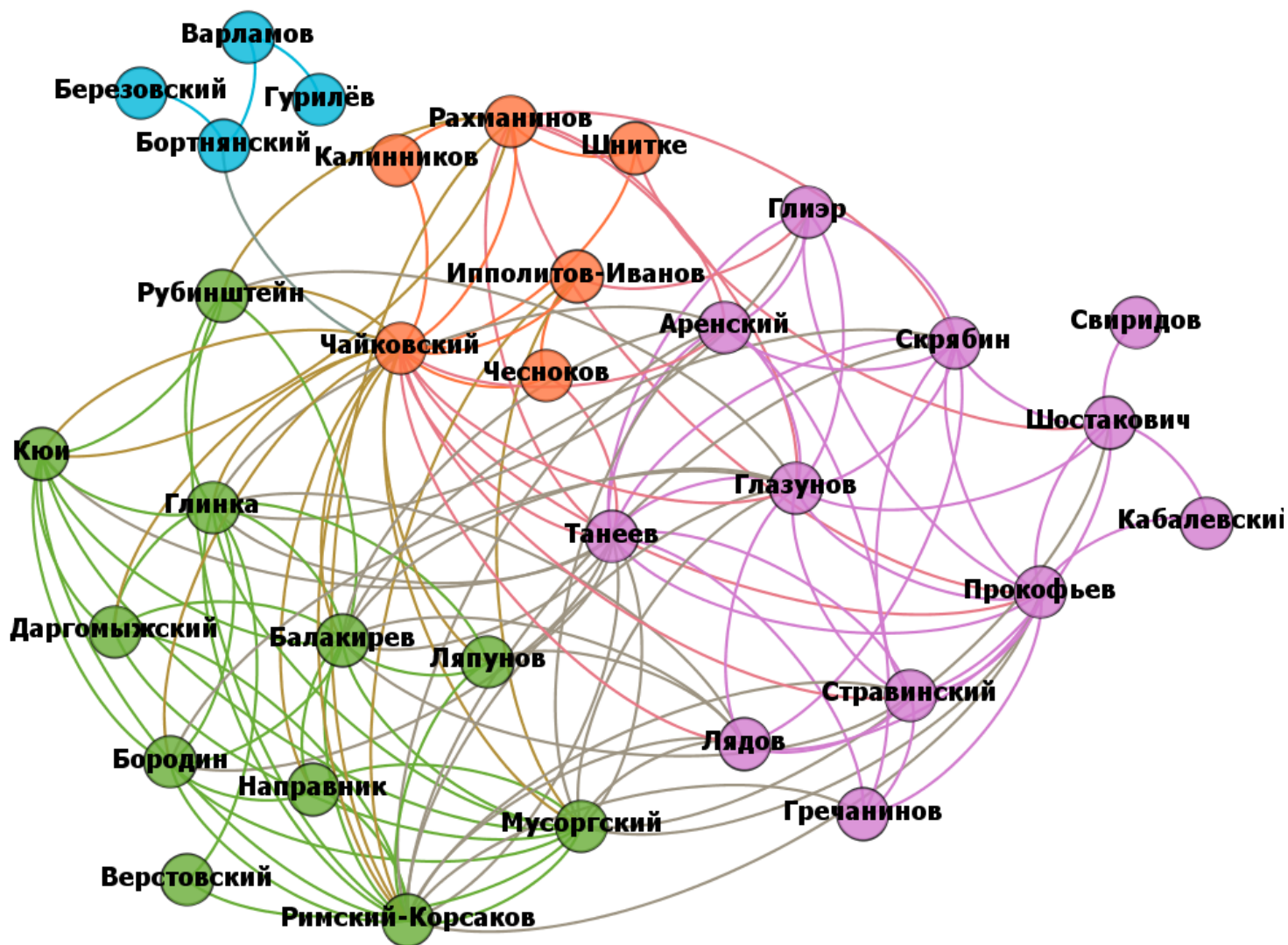
rank	ru	en
1	Римский-Корсаков	Чайковский
2	Рахманинов	Римский-Корсаков
3	Глинка	Танеев
4	Балакирев	Балакирев
5	Мусоргский	Мусоргский

Римский-Корсаков потерял пальму первенства из-за аномально цитируемого Чайковского.

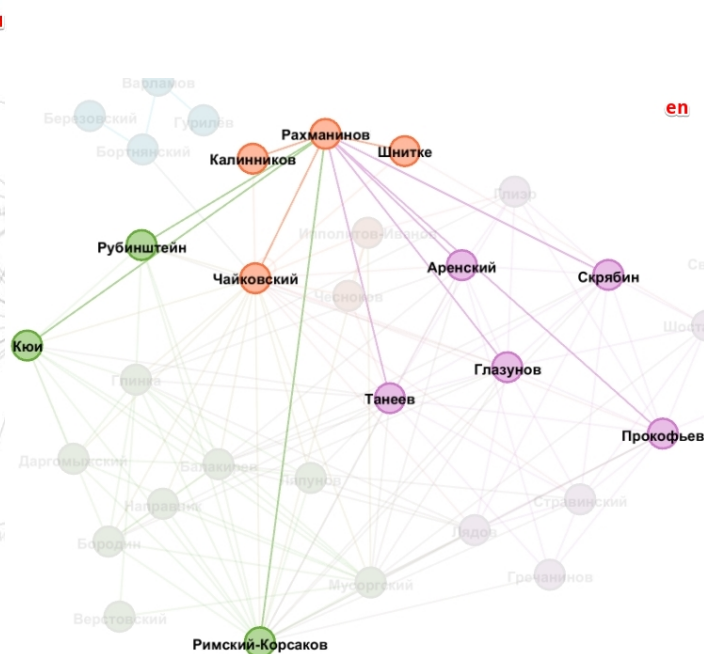
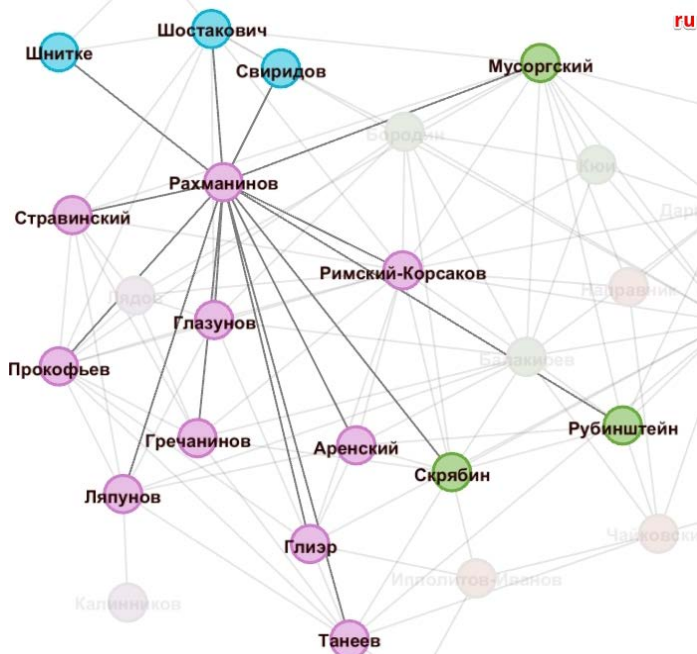
Русская википедия разбитая на сообщества.



Английская версия Википедии разбитая на сообщества



По классам modularity мы видим множество отличий. Посмотрим, например на Рахманинова, который в ru был в топе, а в en уже оказался на 9-ом.



Заметно, что Рахманинов в английской версии перешел в иную группу, при этом не потеряв с некоторыми из них связей: Аренский, Танеев, Прокофьев. Все еще имеет связь с Рубинштейном, Шнитке и т.д., но потерял ее в английской вики с множеством из них: Шостакович, Свиридов, Скрябин, Гречанинов, Ляпунов, Стравинский, зато приобрел Кюи.

Чайковский, возможно, имеет особое значение в западной культуре, ставя его в пример, как «традиционного» русского композитора, своего рода Пушкина в музыке. Однако интересно, почему Рахманинов, оказался так сильно ниже, чем ожидалось. Казалось бы не менее известный русский композитор, чем Римский-Корсаков? Последний, впрочем, остался на своей позиции. Интересно и то, что Танеев выдался в лидеры, согласно Википедии.