Отчёт по курсовому проекту
на тему

«Исследование сочетаемости русских имен прилагательных методами анализа комплексных сетей»

Рыжова Д.А.

1. Характеристики исходного графа

Исследуемая нами лингвистическая сеть строилась на основе подкорпуса Национального корпуса русского языка (SynTagRus), размеченного морфологически и синтаксически.

Узлы графа – 99 наиболее частотных (по словарю Шаров, Ляшевская 2009) качественных прилагательных и связанные с ними в рамках данного корпуса текстов существительные. И прилагательные, и существительные лемматизированы. Рёбра соединяют признаковые и предметные лексемы, связанные между собой атрибутивным синтаксическим отношением.

Граф невзвешенный, двудольный (прилагательные никогда не связываются с прилагательными, существительные – с существительными).

В исходном графе было несколько случаев омонимии существительных и прилагательных, что нарушало двудольность графа (ср. отдал левому **крайнему**, слабые **лёгкие**), но всех этих ситуациях омонимия была искусственно разрешена (узлы получили другие названия).

Количество узлов: 3151 (из них прилагательных: 100, существительных: 3051)
Количество рёбер: 6518

Плотность графа: 0.000656682433896
Распределение степеней узлов: N = 3151, mean +- sd: 4.1371 +- 14.4157

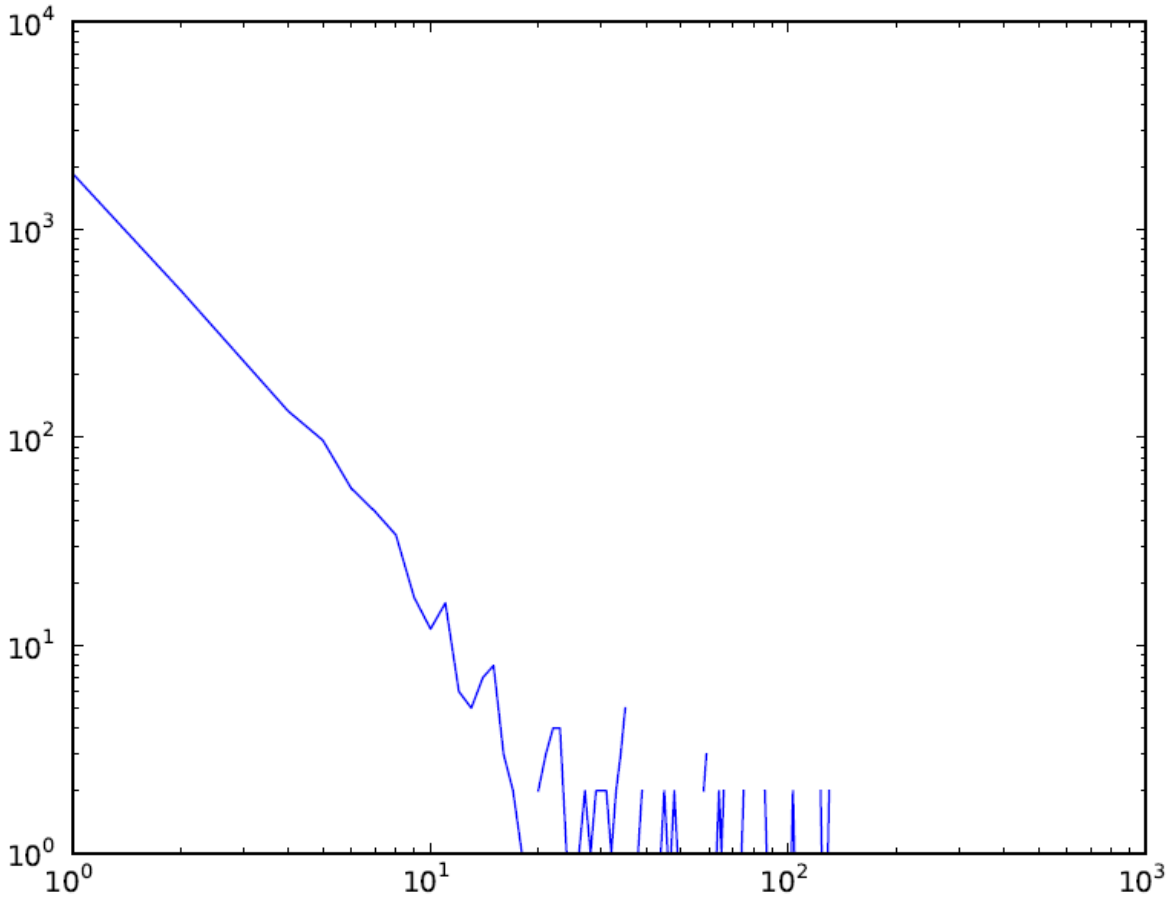

Рис. 1. График распределения степеней узлов

Экспонента распределения alpha:

Exponent (alpha) = 1.901580

p-value = 0.036566

Весь граф представляет собой единую связанную компоненту с диаметром 8 и средней длиной пути 3.93.

Заметим, что наша предыдущая реализация подобного синтаксического графа, узлами которого также были прилагательные и существительные (назовём её граф 1), сильно отличалась от настоящей (граф 2) по числу связанных компонент, по величине диметра и по средней длине пути, ср.:

	ГРАФ 1	**ГРАФ 2**
Связанные компоненты	1 гигантская и много маленьких	1 гигантская
Диаметр	15	8
Средняя длина пути	4,85	3,93

Столь серьёзные количественные изменения связаны, на наш взгляд, с двумя обстоятельствами:

1) с ограничением на частотность прилагательных;
2) с увеличением объёма корпуса.

Высокая частотность слова в языке коррелирует с широкой сочетаемостью, что объясняет большую степень связанности нашего графа и отсутствие «выбросов» - длинных цепочек узлов, из-за которых диаметр оказывается очень большим при достаточно малой средней длине пути. Иными словами, среди частотных прилагательных не найдётся таких, у которых степень узла будет меньше или равна

двум, а существительные не могут быть связаны между собой и, следовательно, не могут образовывать длинных неразветвляющихся цепочек.

Больший объём корпуса позволяет собрать больше статистики, т.е. зафиксировать больше возможных связей, что также увеличивает плотность графа и сокращает среднюю длину пути.

Транзитивность: 0 (в связи с двудольностью графа)
Асортативность: -0.52

2. Центральные узлы

Degree centrality		**Closeness centrality**	
прилагательные	**существительные**	**прилагательные**	**существительные**
большой	человек	отдельный	место
полный	проблема	реальный	проблема
огромный	сила	собственный	машина
крупный	глаз	большой	работа
известный	путь	хороший	дело
собственный	жизнь	последний	жизнь
последний	место	крупный	человек
высокий	деньги	высокий	сила
реальный	работа	огромный	деньги
небольшой	лицо	старый	лицо

Betweenness centrality	**PageRank centrality**
маленький	старый
настоящий	настоящий
старый	огромный
известный	известный
высокий	высокий
крупный	крупный
полный	полный
собственный	собственный
последний	последний
большой	большой

На наш взгляд, очень показательно, что по всем метрикам центральными оказываются, прежде всего, прилагательные со значением общего размера: *большой*, *маленький*, *крупный*, *огромный*. Это означает, что сочетаемость таких слов наименее избирательна. Так, например, признак *широкий*, также выражающий значение размера, но не общего, а более специфического (как правило, только в горизонтальном измерении и только плоских предметов), имеет достаточно серьёзные ограничения на сочетаемость: *широкий пояс*, *широкая доска*, *широкая площадка*, но не **широкий кактус*, **широкая кастрюля* и т.п. Прилагательное *большой* выражает менее сложное значение, поэтому и сочетаемость его значительно шире.

Любопытно, что при этом центральным по всем метрикам оказывается и прилагательное *высокий*, которое, казалось бы, стоит в одном ряду с признаком *широкий* и тоже выражает сложное значение размера (величину вертикально ориентированных предметов, как правило, закрепленных снизу: *дерево*, *забор*, *здание*,

но не **река*, **верёвка*, **рука*). Вероятнее всего, центральность признака *высокий* обеспечивается его распространёнными (почти клишированными) употреблениями в переносных значениях: ср. *высокий уровень*, *высокое положение в обществе* и т.п.

3. Сравнение со случайным графом

Сгенерированный случайный граф Erdos-Renyi отличается от лингвистического:

	случайный	лингвистический
средняя степень узлов	4.1371 +- 2.0468	4.1371 +- 14.4157
средняя длина пути	5.77	3.93
транзитивность	0.0019	0

Случайный граф характеризуется меньшим разбросом в значениях степеней узлов, ненулевой транзитивностью и более высоким значением средней длины пути. Средняя длина пути в случайном графе хорошо моделирует свойство малого мира (правило шести рукопожатий) и ещё раз демонстрирует, что сконструированный нами лингвистический граф моделирует «сверхмалый» мир – вероятнее всего, из-за ограничения на частотность прилагательных.

4. Core structure

Максимально возможное значение core для нашего графа – 8. Этот ядерный подграф, узлы которого связаны между собой наиболее тесным образом, включает в себя 771 ребро из 124 узла:

- Прилагательные:

Простой, важный, прямой, быстрый, молодой, полный, плохой, чистый, маленький, целый, средний, чужой, последний, старый, нормальный, мелкий, значительный, крупный, высокий, тяжёлый, большой, небольшой, огромный, мировой, внутренний, собственный, хороший, мягкий, близкий…

- Существительные:

Дело, решение, система, тело, человек, задача, путь, вопрос, система, глаз, вещь, жизнь, успех, эффект, метод, число, работа, место, исследование, ресурс, значение, процесс, событие, организация, проблема, проект, мера, программа, форма, результат, время, слово, день, женщина…

Характерно, что в ядро попадают прилагательные, выделявшиеся метриками центральности.

5. Сетевые сообщества

1) edge betweenness community

Метрика edge betweenness разбивает граф на слишком большое число сообществ:

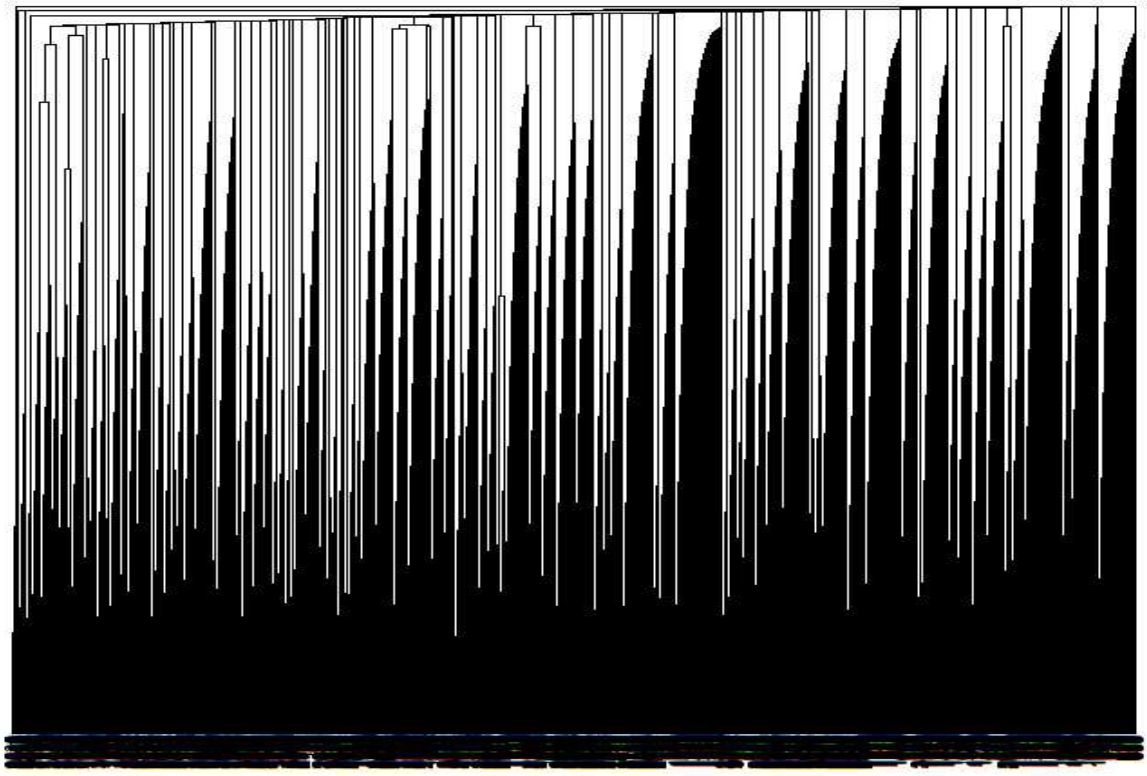

Рис. 2. Edge betweenness communities: дендрограмма

2) walktrap

Аналогично слишком дробное разделение получается по метрике waltrap (3051 узел делятся на 3050 сообществ):

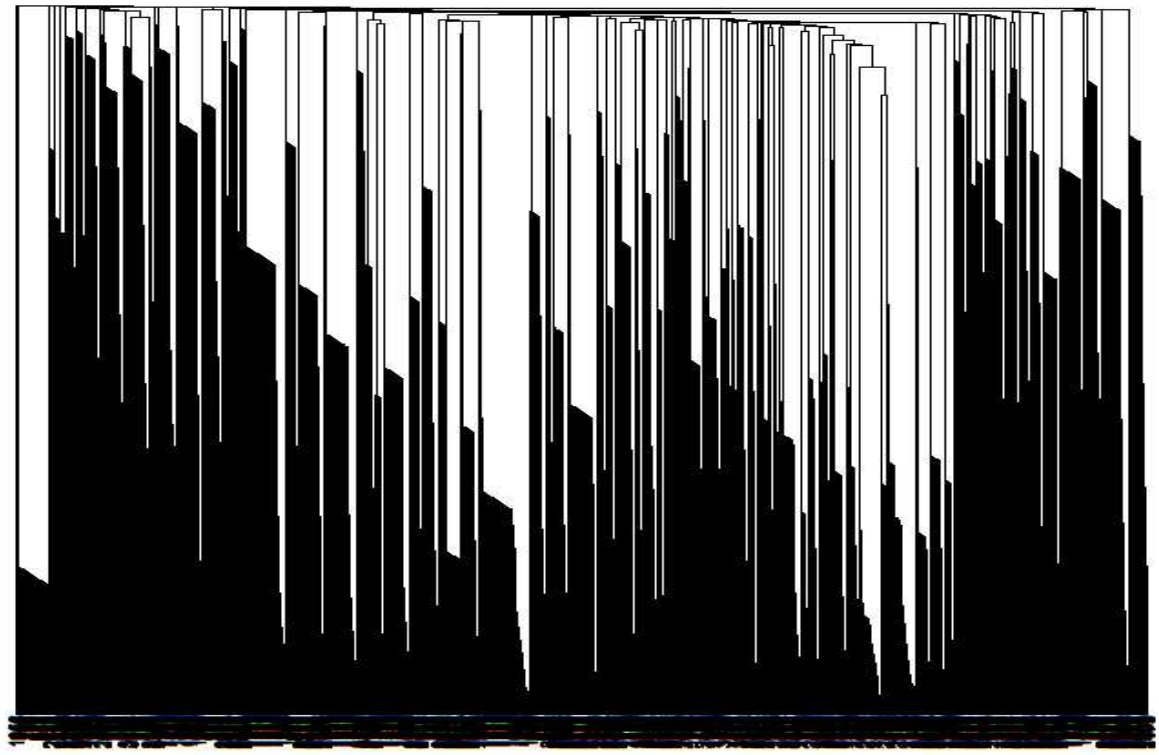

Рис. 3. Walktrap communities: дендрограмма

3) eigenvector

Метрика eigenvector, напротив, выделяет обозримое число сообществ (22), многие из которых интуитивно кажутся содержательными. Ср. один из кластеров, в который попало четыре прилагательных: *тёмный*, *зелёный*, *жёлтый* и *маленький*. Видно, что 3 прилагательных из четырёх – цветообозначения, а прилагательное *маленький* сочетается с особым типом существительных: с существительными, оформленными уменьшительно-ласкательными суффиксами (*ключик*, *человечек*, *комарик*, *корытце*, *рамка*…).

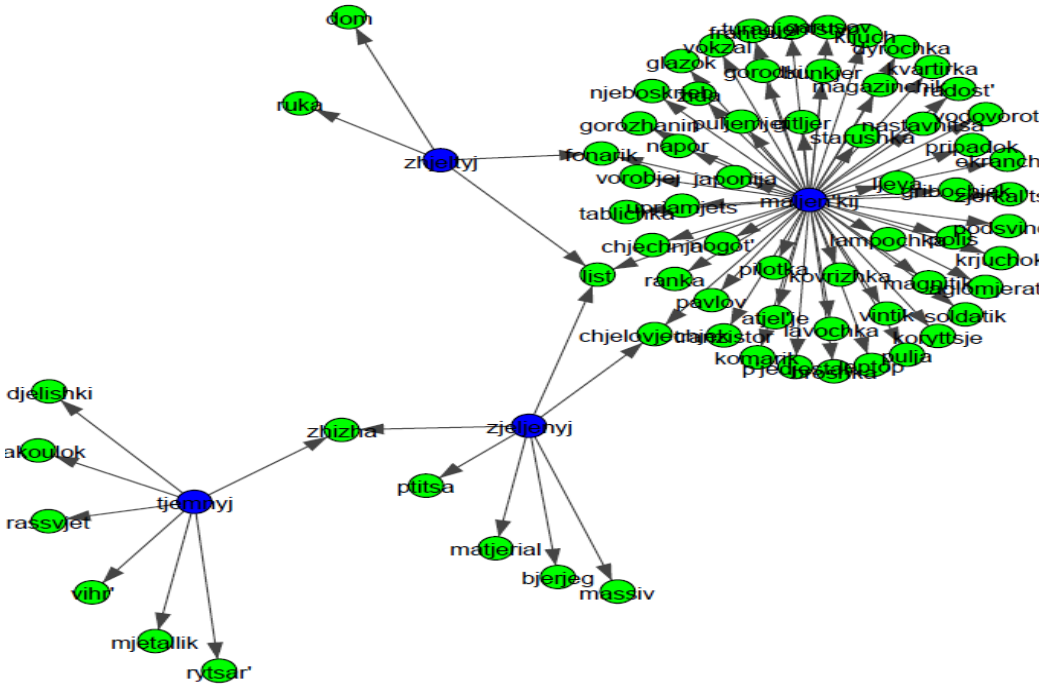

Рис. 3. Сообщество, выделенное по метрике eigenvector. Синим цветом обозначены прилагательные, зелёным - существительные

6. Триады и диады

Диады: 0 mutual, 6518 asymmetric[1], 4956307 null dyads

Триады:
021D – binary out-tree: 335534
021U – binary in-tree: 12220
Отсутствие тернарных связей объясняется двудольностью графа.

7. Структурная схожесть узлов

В нашем случае наиболее осмысленно было использовать метрику similarity jaccard, подразумевающее отношение общих для двух узлов соседей к объединению всех соседей рассматриваемых узлов.

Мы отдельно вычисляли близость для узлов-прилагательных и отдельно – для узлов-существительных.

Среди прилагательных большую близость демонстрировали антонимы. Так, например, прилагательные *тяжёлый* и *лёгкий* больше похоже друг на друга (значение метрики 0,04), чем прилагательные *тяжёлый* и *большой* (0,028).

[1] В связи с тем, что граф был представлен списком рёбер, каждое ребро – парой узлов, а в каждой паре на первом месте стояло прилагательное, а на втором – существительное, граф воспринимался как направленный. Для нас в большинстве случаев направленность графа не имела значения.

Максимальное сходство (0,2) - у «парных» прилагательных *правый* и *левый*, *верхний* и *нижний*.

Среди существительных максимально похожими (значение метрики 0,6) оказались слова *тезис* и *владелец*. Во многом их близость объясняется тем, что они оба характеризуются не очень большим числом соседей (что обеспечивает им заведомо небольшую разницу в контекстах) и при этом оба соседствуют с прилагательными *простой*, *полезный*, *известный*.

8. Выводы и визуализации

Из-за того, что в граф включались только частотные прилагательные, которые обычно, как уже было сказано выше, характеризуются широкой сочетаемостью, граф получился очень плотным и сложно визуализируемым (ср. рис. 4).

С другой стороны, граф можно разделить на фрагменты (например, grouping – natural clusters в среде yEd), по которым видно, как именно наш граф устроен: основные, структурообразующие узлы – прилагательные, вокруг которых собирается пучок существительных (ср. рис. 5). Некоторые существительные попадают сразу в несколько пучков (ср. рис. 6), из-за чего граф становится плотнее (особенно это характерно для ядра).

Однако очень важно, что в графе не все узлы связаны со всеми. Это говорит о том, что у каждого прилагательного есть свои определённые ограничения на сочетаемость. Например, несмотря на то, что цвет есть, в принципе, у всех физических предметов, далеко не все предметные имена сочетаются с именами цветообозначений (ср. неестественность словосочетаний **чёрный комар* или **голубая речка*).

Для того, чтобы подтвердить тезис об избирательности сочетаемости прилагательных, можно было провести ещё один, более чистый эксперимент: взять в качестве узлов исключительно прилагательные со значениями физических признаков и существительные, обозначающие только физические предметы. При этом узлов в графе должно быть как можно меньше, чтобы визуализации были наглядны. Например, можно взять группу прилагательных размера и сочетающиеся с ними предметные имена и проверить, будет ли разница в сочетаемости этих признаковых слов улавливаться методами анализа комплексных сетей.

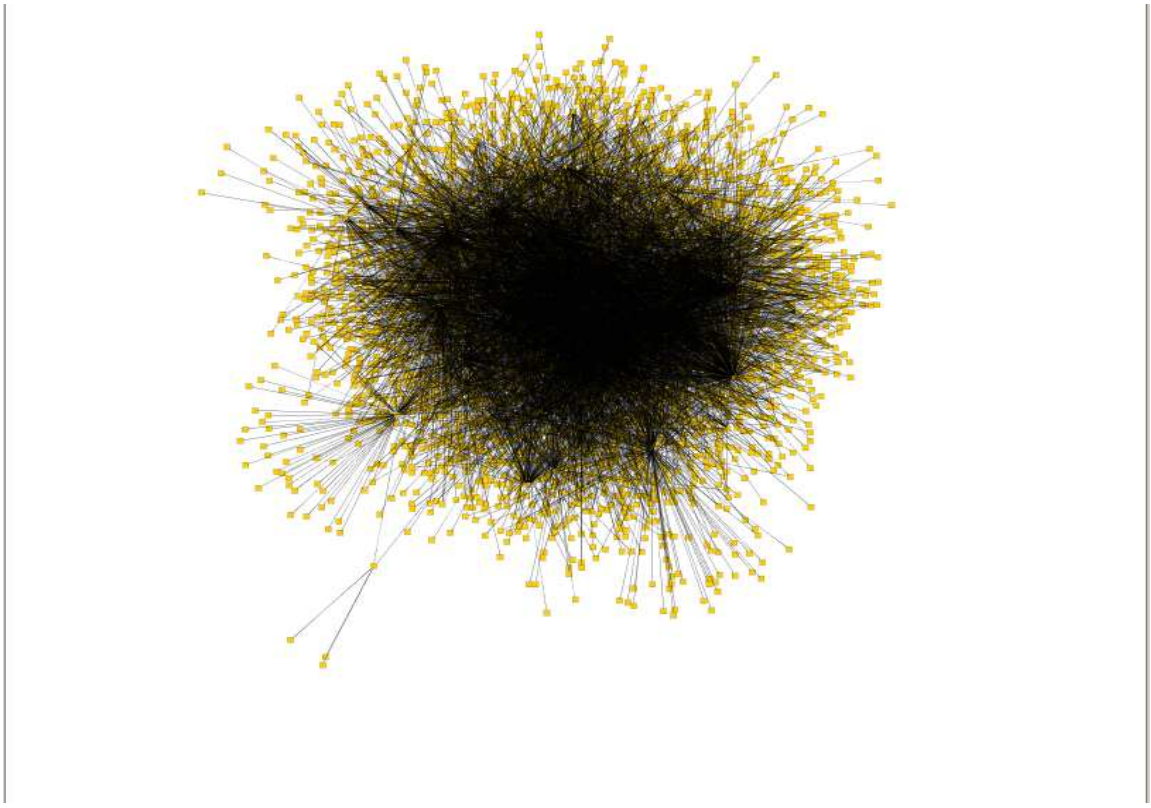

Рис. 4. Визуализация графа

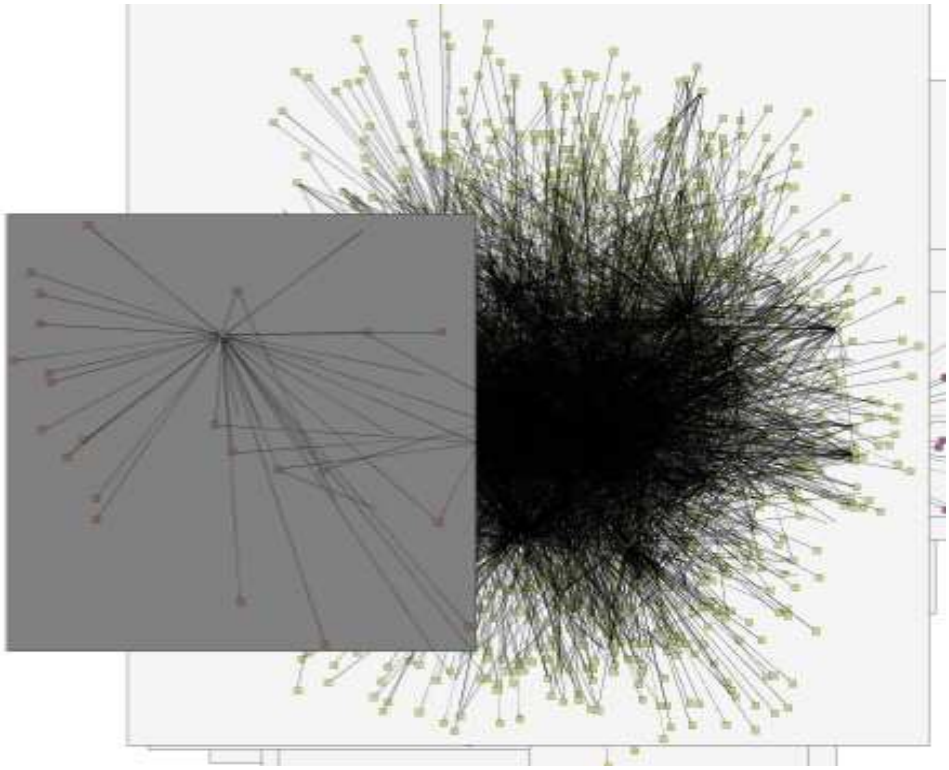

Рис. 5. Natural clusters

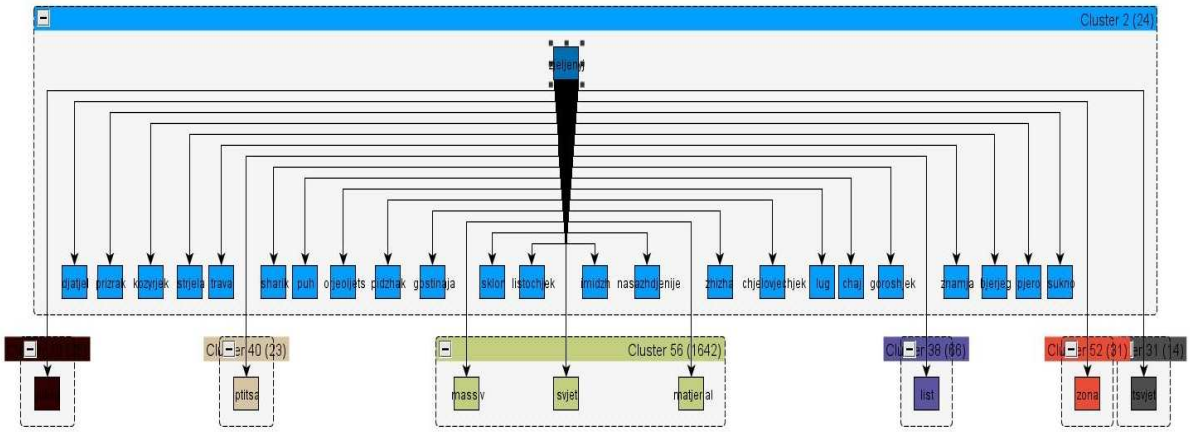

Рис. 6. Соседи узла зелёный (сам узел находится наверху, голубые узлы попадают в кластер с центром зелёный, узлы других цветов – в другие кластеры).