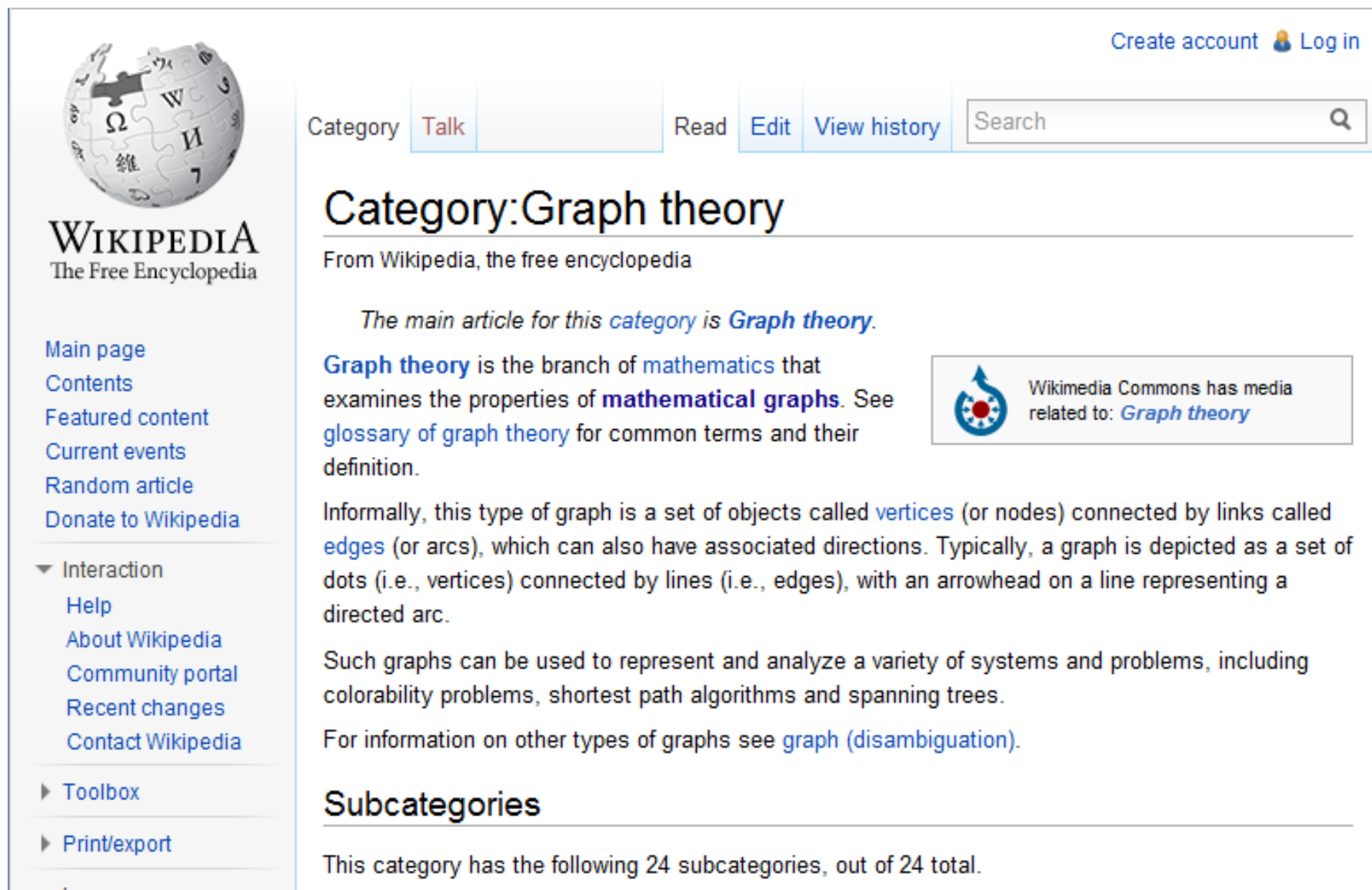


Исследование графа категорий английской версии Wikipedia

Александр Шкотин
ashkotin@acm.org

Государственный Геологический Музей РАН
Москва

Пример категорной страницы-1



The screenshot shows the Wikipedia page for the category "Graph theory". At the top right, there are links for "Create account" and "Log in". Below these are navigation tabs: "Category", "Talk", "Read", "Edit", and "View history". A search box is located to the right of these tabs. The main heading is "Category:Graph theory", followed by the text "From Wikipedia, the free encyclopedia". A note states: "The main article for this category is [Graph theory](#)." The main text begins with "Graph theory is the branch of mathematics that examines the properties of **mathematical graphs**. See [glossary of graph theory](#) for common terms and their definition." A box on the right indicates "Wikimedia Commons has media related to: [Graph theory](#)". The text continues: "Informally, this type of graph is a set of objects called **vertices** (or nodes) connected by links called **edges** (or arcs), which can also have associated directions. Typically, a graph is depicted as a set of dots (i.e., vertices) connected by lines (i.e., edges), with an arrowhead on a line representing a directed arc." It then says: "Such graphs can be used to represent and analyze a variety of systems and problems, including colorability problems, shortest path algorithms and spanning trees." and "For information on other types of graphs see [graph \(disambiguation\)](#)." Below this is a section titled "Subcategories" with the text: "This category has the following 24 subcategories, out of 24 total." The left sidebar contains the Wikipedia logo and navigation links: "Main page", "Contents", "Featured content", "Current events", "Random article", "Donate to Wikipedia", "Interaction" (with sub-links: "Help", "About Wikipedia", "Community portal", "Recent changes", "Contact Wikipedia"), "Toolbox", and "Print/export".

Пример категорной страницы-2

The image shows a screenshot of a web browser displaying the Wikipedia category page for "Graph theory". The browser's address bar shows the URL "en.wikipedia.org/wiki/Category:Graph_theory". The page content includes a sidebar with navigation options like "Contact Wikipedia", "Toolbox", "Print/export", and "Languages". The main content area features a note about disambiguation, a "Subcategories" section listing 24 subcategories grouped by letter (A, C, D, E, G, I, M, N, O, R, T), and a section for "Pages in category 'Graph theory'".

Category:Graph theory - Wiki x

en.wikipedia.org/wiki/Category:Graph_theory

Contact Wikipedia

- ▶ Toolbox
- ▶ Print/export
- ▼ Languages
 -
 - العربية
 - Aragonés
 - Беларуская
 - Boarisch
 - Bosanski
 - Български
 - Català
 - Česky
 - Dansk
 - Deutsch
 - Eesti
 - Ελληνικά
 - Español
 - Esperanto
 - Euskara
 - فارسی
 - Français
 - 한국어
 - Ido

For information on other types of graphs see [graph \(disambiguation\)](#).

Subcategories

This category has the following 24 subcategories, out of 24 total.

A <ul style="list-style-type: none">▶ Algebraic graph theory (1 C, 42 P)	G <ul style="list-style-type: none">▶ Geometric graph theory (2 C, 14 P)▶ Graph description languages (9 P)▶ Graph minor theory (20 P)▶ Graphs (6 C, 8 P)	O <ul style="list-style-type: none">▶ Graph theory objects (3 C, 37 P)▶ Graph operations (1 C, 21 P)
C <ul style="list-style-type: none">▶ Graph coloring (1 C, 50 P)▶ Computational problems in graph theory (2 C, 45 P)▶ Graph connectivity (32 P)	I <ul style="list-style-type: none">▶ Graph invariants (1 C, 66 P)	R <ul style="list-style-type: none">▶ Random graphs (13 P)▶ Graph rewriting (7 P)
D <ul style="list-style-type: none">▶ Graph data structures (27 P)▶ Graph databases (7 P)	M <ul style="list-style-type: none">▶ Matching (30 P)▶ Mathematical chemistry (20 P)	T <ul style="list-style-type: none">▶ Theorems in graph theory (36 P)▶ Graph theorists (1 C, 94 P)▶ Topological graph theory (1 C, 28 P)▶ Trees (data structures) (8 C, 97 P)
E <ul style="list-style-type: none">▶ Graph enumeration (4 P)▶ Extremal graph theory (8 P)	N <ul style="list-style-type: none">▶ Network theory (3 C, 43 P)	

Pages in category "Graph theory"

Пример категорной страницы-3

Category:Graph theory - Wiki

en.wikipedia.org/wiki/Category:Graph_theory

Network theory (3 C, 43 P)

Pages in category "Graph theory"

The following 100 pages are in this category, out of 100 total. This list may not reflect recent changes ([learn more](#)).

- Glossary of graph theory
- Graph theory
- List of graph theory topics

A

- Aanderaa–Karp–Rosenberg conjecture
- Aczel's anti-foundation axiom
- Adjacent
- Alpha centrality
- Assortative mixing
- Assortativity

B

- Baker's technique
- Betweenness centrality
- Bicircular matroid
- Bondy's theorem

E cont.

- Edge-graceful labeling
- Eigenvalues and eigenvectors of the second derivative
- Erdős–Burr conjecture
- Erdős–Gyárfás conjecture
- Erdős–Gallai theorem
- Evolutionary graph theory
- Expander mixing lemma

F

- Forbidden graph characterization
- Frequency partition of a graph
- Friendship paradox

G

M

- Markov chain
- Mathematical chemistry
- Maximum common edge subgraph problem
- Modular decomposition
- Multi-level technique
- Multi-trials technique

N

- Network theory

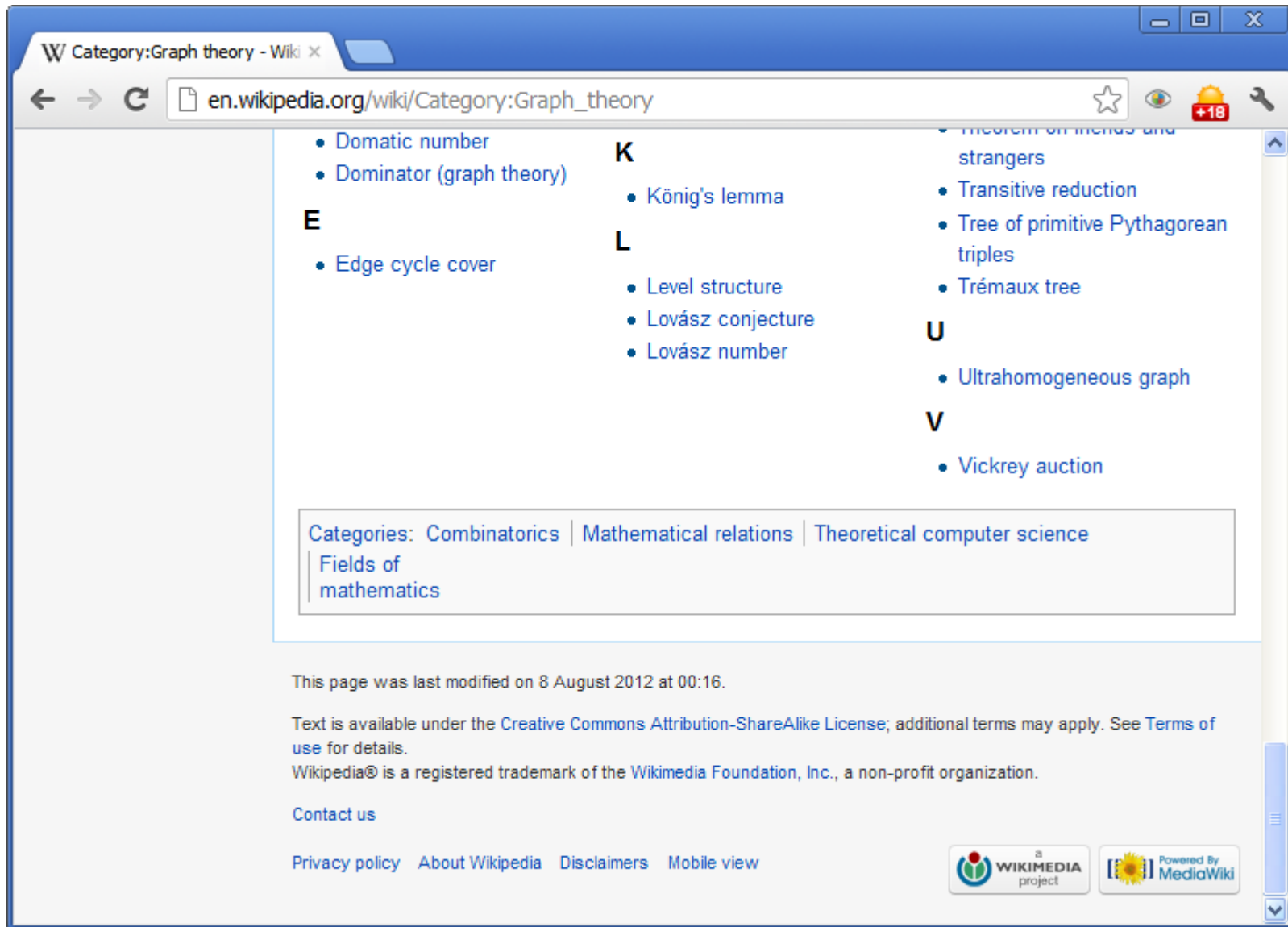
P

- Peripheral cycle
- Power graph analysis

R

- Ramsey's theorem
- Random graph

Пример категорной страницы-4



Граф категорий Википедии (ГКВ)

ГКВ есть оргграф каждый узел которого взаимно-однозначно соответствует категорной странице и помечен её номером.

Дуга из узла N_1 в узел N_2 идёт тогда и только тогда, когда страница с номером N_1 есть под-категория страницы с номером N_2 .

Исходные данные («дамп») [Антон Коршунов, ИСП РАН]:

- $N_1 N_2$ flag
- N title

flag: 0 - N_1 - простая страница, 1 - N_1 - категорная.

ГКВ. Характеристики

Количество узлов - 593796 (изолированных - 26272).

Количество стрелок - 1221133. Петель нет.

Количество связных компонент - 1987. Количество узлов в 10-и самых больших компонентах: 561636, 210, 36, 29, 27, 20, 19, 19, 17, 16.

Источников - 345597. Промежуточных узлов - 210160. Стоков - 11767. Это в основном «точки роста».

"Category:World War II" - 85 исходящих.

"Category:Albums by artist" - 12625 входящих.

Путь: Category:Anastacia songs → Category:Music, **154** узла.

Самый длинный - **294** узла.

ГКВ в целом

Есть ор-циклы [1], с.9 - чего быть не должно.

Ядро - объединение ор-циклов графа и ор-путей между циклами.

Мантия - дополнительная часть графа.

Связующие стрелки - между мантией и ядром. Часть идёт из ядра в мантию (591), а часть - из мантии в ядро (210514).

Ядро получается итеративным удалением источников и стоков.

Ядро ГКВ

Стрелок в ядре - 38538. Узлов - 13545.

Связные компоненты: одна большая - 13507 узлов. И ещё 19 пар узлов. Таким образом большинство связанных компонент ГКВ ядер не имеет.

Чисто ядерная связная компонента ГКВ:

Category:Wikipedia sockpuppets of ShantanuSingh198

Category:Suspected Wikipedia sockpuppets of ShantanuSingh19

Сильно связанные компоненты ядра-1

В ядре нас интересует зацикливание отношения под-категория - категория. Два подхода:

- общий - применить алгоритм поиска сильно связанных компонент (ССК);
- частный - найти так называемые "линзы" - два узла ссылающиеся друг на друга (как под-категория - категория).

Математически цикл утверждает эквивалентность соответствующих терминов, т.е. синонимию, что в принципе возможно.

Чтобы получить состав сильно связанных компонент ядра была использована программа Rajek [2].

Сильно связанные компоненты ядра-2

Количество ССК - 457.

Узлов не входящих в ССК — 7646.

Гигантская ССК - 3967 узлов.

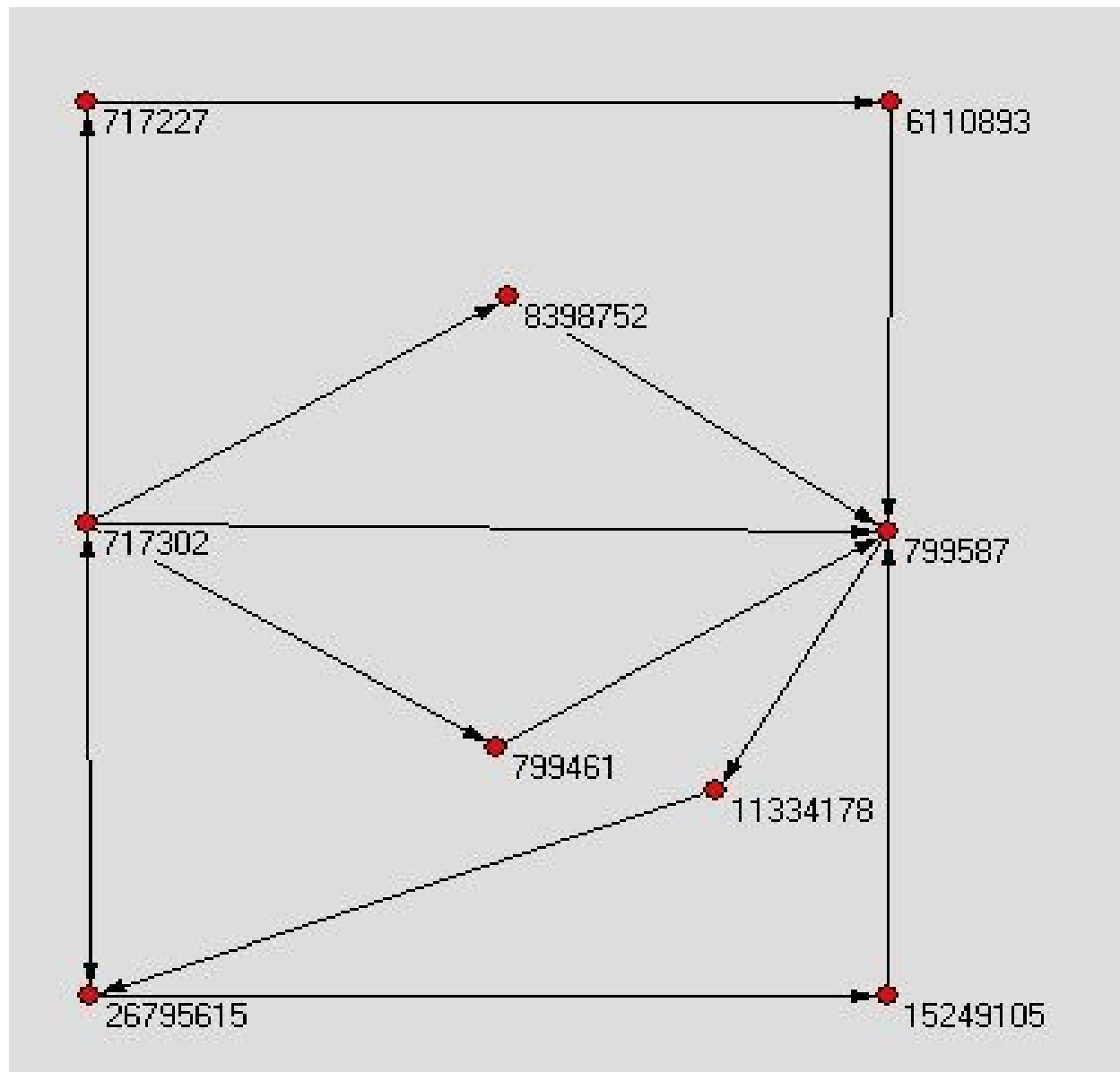
Количество узлов в следующих по величине ССК:

96, 95, 68, 52(7), 51(3), 18.

Линз - 1269.

Рассмотрим компоненту №41 у которой всего 9 узлов.

ССК 41. Рисунок графа компоненты



ССК 41. Заголовки узлов

717227	Category:Orthodox rabbis
717302	Category:Talmud rabbis
799461	Category:Mishnah
799587	Category:Talmud
6110893	Category:Talmudists
8398752	Category:Talmud people
11334178	Category:Rabbinic literature
15249105	Category:Talmud concepts and terminology
26795615	Category:Chazal

Мантия ГКВ отдельно

Изолировавшихся источников - 14421, стоков - 60.

Не изолированных стоков - 11707.

Ложных стоков - 18157.

Максимальная высота у ложного стока «Category:Creation myths».

Максимальная высота настоящего стока - 24.

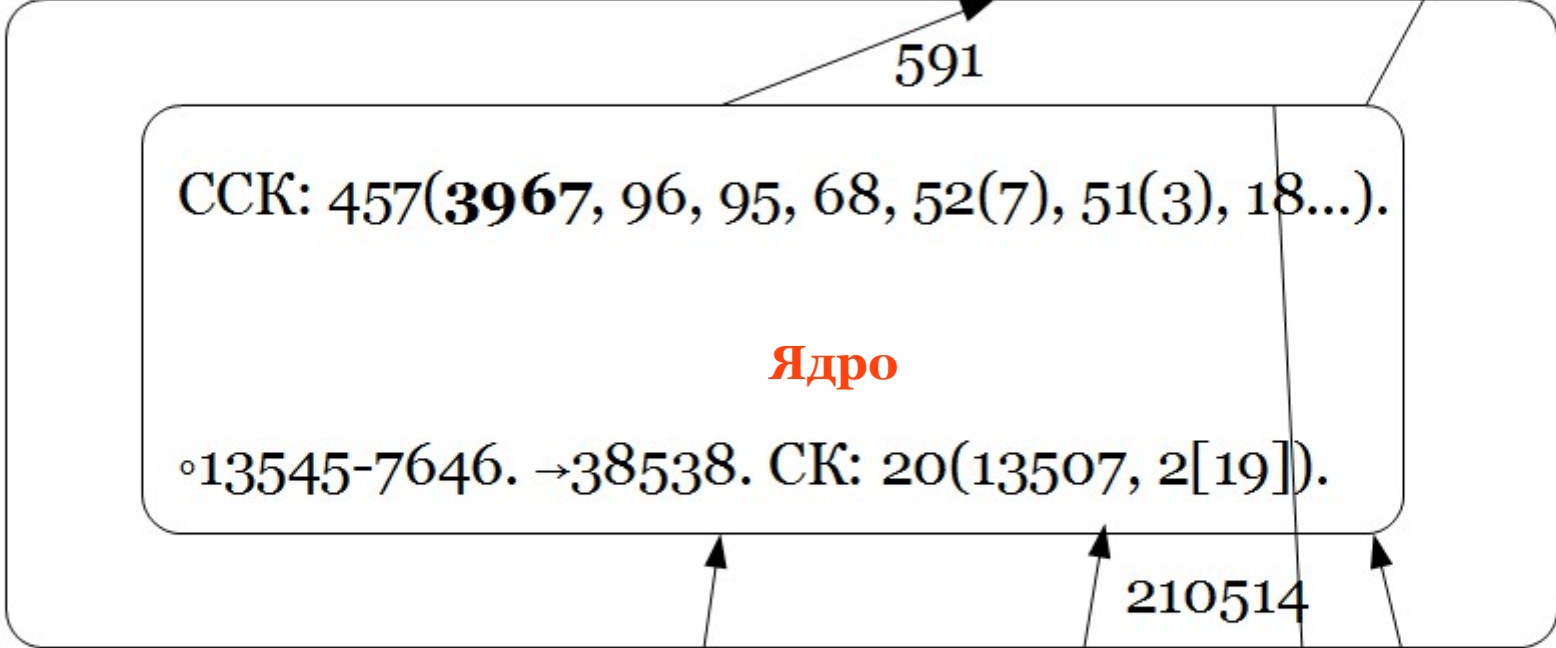
В таблице: level = NULL - количество изолированных узлов мантии, level = 0 - количество узлов в ядре.

level	count
NULL	14481
28	1
27	2
26	3
25	3
24	5
23	7
22	12
21	16
20	20
19	30
18	50
17	57
16	71
15	100
14	149
13	226
12	425
11	697
10	1187
9	1915
8	3103
7	4858
6	7754
5	13019
4	23302
3	45323
2	105958
1	331205
0	13545

11707. T24

Изолировавшиеся

60



ССК: 457(3967, 96, 95, 68, 52(7), 51(3), 18...).

Ядро

◦13545-7646. →38538. СК: 20(13507, 2[19]).

18157. (T)28

-29

14421

◦593796-26272. →1221133. СК: 1987(561636, 210...2).

level	count
NULL	14481
28	1
27	2
26	3
25	3
24	5
23	7
22	12
21	16
20	20
19	30
18	50
17	57
16	71
15	100
14	149
13	226
12	425
11	697
10	1187
9	1915
8	3103
7	4858
6	7754
5	13019
4	23302
3	45323
2	105958
1	331205
0	13545

Обсуждение

ГКВ должен быть ациклическим графом. Таким образом аномалии значительны.

Можно создать средства уведомления о логическом противоречии. Для начала можно разобраться с 1269 линзами.

Интересно исследовать длинные пути.

Многие отношения между категориями попавшие в sub-category of следует перенести в See also.

Важная задача - получить и проанализировать русский ГКВ.

Литература

1. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Soren Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann. DBpedia - A Crystallization Point for the Web of Data. May 25 2009.
2. Batagelj V., Mrvar A. Pajek reference manual. Ljubljana, April 16, 2012.
3. Исследовательский отчёт "Исследование графа категорий английской версии Wikipedia": [текст](#).

Спасибо за внимание!

Александр Шкотин

ashkotin@acm.org