Факультет компьютерных наук

Прикладная математика и информатика

Москва 2023

Сетевые модели исследований в области болезни Паркинсона

Network Models of Studies on Parkinson Disease

Работу выполнила:
Степочкина Анна, БПМИ194

Научный руководитель:
Алескеров Ф. Т.

Тематическая область

Болезнь Паркинсона (БП) - это нейродегенеративное, прогрессирующее заболевание, в основном характерное для людей старшего возраста.

В США в 2020 году 930 тысяч человек живут с БП, а к 2030 это число возрастет до 1.2 миллиона, каждый год болезнь диагностируется у 60 тысяч людей.

Прямые и косвенные затраты на БП в США составляют 52 миллиарда долларов ежегодно, на лекарства - $2500, а операции достигают стоимости в 100 тысяч долларов на одного человека.

Цели

- Проанализировать основных участников в области исследований БП
- Провести апробацию новых методов сетевого анализа
- Разработать подход к анализу научной области с использованием сетевых моделей

Этапы работы

1. Обзор литературы и источников
2. Сбор данных из базы Microsoft Academic
3. Обработка данных
4. Предварительный анализ
5. Построение сети цитирования для статей и аффилиаций
6. Расчет индексов центральности по сетям
7. Выделение топ-10 вершин по индексам, их сопоставление
8. Анализ динамических изменений в индексах по годам
9. Анализ устойчивости сети
10. Объединение методов в комплексную методологию
11. Основные выводы и направление дальнейших исследований
12. Подготовка статьи по данной теме

Обзор литературы

Анализ публикаций по БП:

- количество цитирований [1],
- индекс Хирша [2],
- анализ статей с числом цитирования большим 400 [3],
- кластерный анализ сетей цитирований по статьям об успешном старении [4] и коронавирусе [5],
- исследование изменения структуры коллаборативной модели в стратегическом менеджменте [6],
- индексы центральности SRIC и LRIC в работе 2018 года [7].

Используя сетевой анализ цитирований, можно выделить ключевые исследования и журналы, в которых они публикуются.

Описание данных

Из базы научных публикаций Microsoft Academic были скачаны статьи 2015-2021 года со словами “parkinson” и “disease” в названии или аннотации.

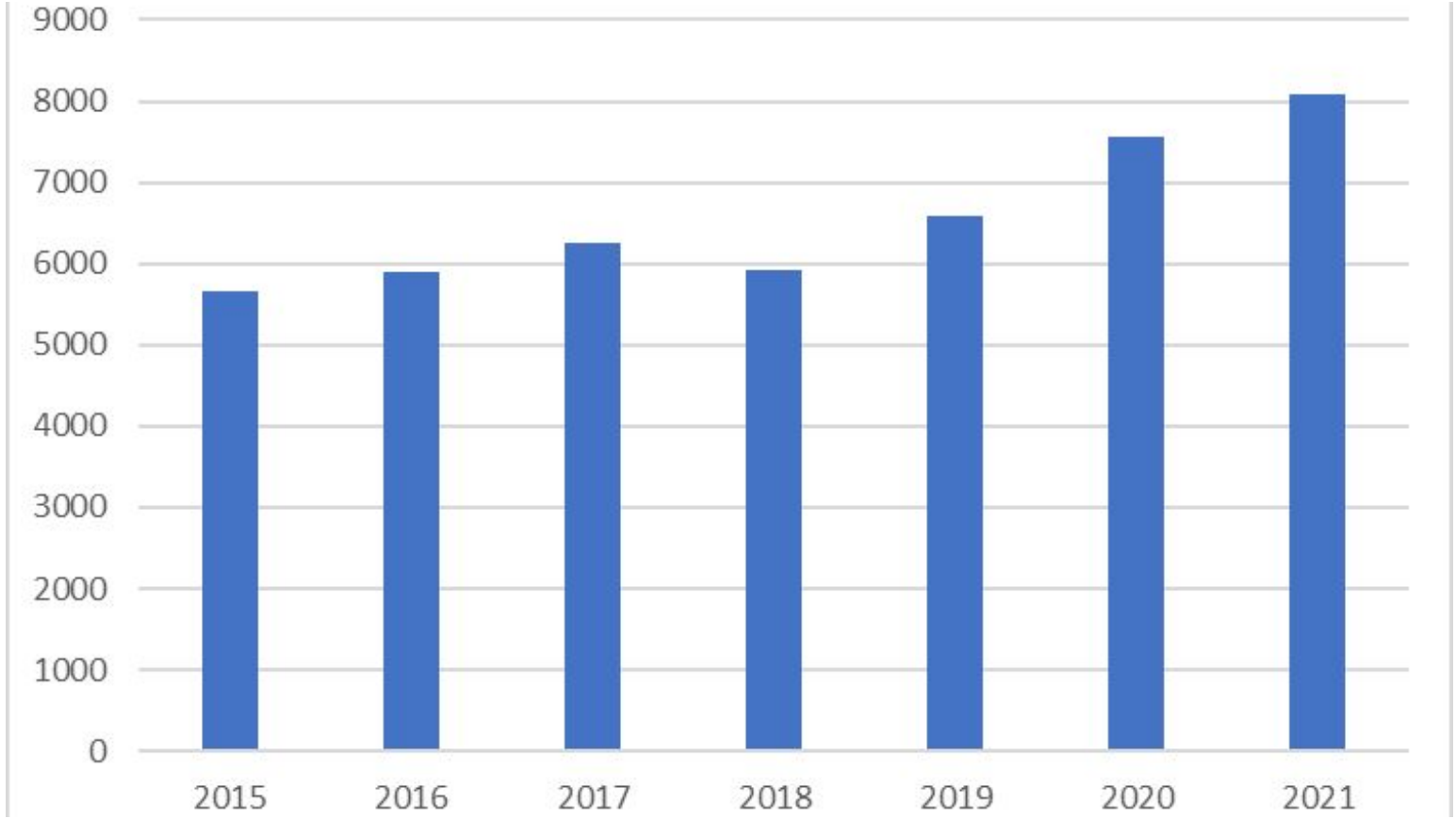
Атрибуты статьи:

- Id – ID публикации
- W, AW – уникальные нормализованные слова (приведенные к единой форме) в заголовке и абстракте
- Y – год публикации
- RId – список ID статей, на которые ссылается публикация
- DOI – цифровой идентификатор объекта
- J.JId – ID журнала

Всего статей скачано: 70119.
Из них

- **10681 без DOI**
- **7315 с DOI и без аннотации**
- **52123 с DOI и аннотацией**
- **45940 с DOI, аннотацией и журналом**

Распределение статей с DOI, аннотацией и журналом по годам



Сеть цитирований по статьям

Сеть цитирований – ориентированный граф, где статьи – вершины, а ребра – цитирования. Если статья ID1 ссылается на статью ID2, то появляется направленное ребро от ID1 к ID2. Y – год статьи ID1.

Количество ребер 2698197, вершин - 836533.

ID1	ID2	weight	Y
2754967293	2339791932	1	2017
2112455323	1920030402	1	2015
2177834950	2123627348	1	2015
2558041282	1934236512	1	2016
2558041282	1947901277	1	2016
2558041282	1564387586	1	2016
2558041282	1903888485	1	2016
2558041282	2338408645	1	2016
2558041282	2508023300	1	2016
2558041282	2097335121	1	2016

Обработка данных

Удалены:

- 115 статей с синдром Вольфа-Паркинсона-Вайта
- Вершины не из первоначального множества (например, если есть ссылка на статью, которая не содержит слов “parkinson” и “disease”)
- Изолированные вершины
- Ребра, в которых год ссылающейся публикации меньше чем год цитируемой (4625 ребер, 105 вершин)

	Количество вершин	Количество ребер
Исходная сеть	780796	2488343
После удаления ребер, где хотя бы одна вершина не входит в исходный набор статей	39825	312044
После удаления неправильных ребер	**39811**	**310829**

Сеть цитирований по статьям

Количество вершин	39811
Количество ребер	310829
Количество компонент связности	94
Размер наибольшей компоненты связности	39618
Плотность графа	0.0002
Минимальное количество цитирований	0
Максимальное количество цитирований	1563
Среднее количество цитирований	7.8

$$density = \frac{m}{n(n-1)}$$

m - количество ребер
n - количество вершин

Индексы центральности

Классические индексы

1. In-degree index – сумма весов входящих ребер

$$x_i = \sum_j w_{ij}$$

2. Eigenvector index – решение уравнения $Ax = \lambda_1 x$, где λ_1 максимальное собственное значение матрицы смежности A

$$x_i = \frac{1}{\lambda_1} \sum_j A_{ij} x_j$$

Классические индексы

3. PageRank centrality – это разновидность центральности по собственному вектору, которая учитывает исходящую степень

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta$$

4. Betweenness centrality – показывает долю кратчайших путей между двумя вершинами, на которых лежит исследуемая вершина

$$x_i = \sum_{kj} \frac{n_{kj}^i}{g_{kj}}$$

Новые индексы

2. Pivotal index (PI) показывает влияние ключевых вершин.

Вершина j_p называется ключевой, если $\sum_{j \in S} w_{ji} \geq q_i$ и $\sum_{j \in S \setminus \{j_p\}} w_{ji} < q_i$.

Значение индекса для критического множества равно количеству ключевых вершин в нем. Для вершины: $PI(i) = \sum_{S} |S| \times PI_i(S)$

Общее влияние: $TI(i) = \frac{1}{3} In - degree(i) + \frac{1}{3} BI(i) + \frac{1}{3} PI(i)$

Новые индексы

S – критическое множество для вершины i, если $S \subseteq V\backslash\{i\}$,

$|S| \leq k, \sum_{j \in S} w_{ji} \geq q_i$, где квота q_i - процент суммы весов входящих ребер, k – количество вершин, которые одновременно могут влиять на узел. В работе k=3. С увеличением квоты уменьшается количество критических множеств.

1. Bundle index (BI) учитывает групповое влияние на вершину

$$BI_i(S) = 1, \text{если} \sum_{j \in S} w_{ji} \geq q_i \qquad BI_i(S) = 0, \text{иначе}$$

$$BI(i) = \sum_{S} BI_i(S)$$

Aleskerov F., Yakuba V. Matrix-vector approach to construct generalized centrality indices in networks. 2020 // SSRN 3597948. Available at: https://ssrn.com/abstract=3597948

Паттерн-анализ

1. Представление данных в виде графика на параллельных координатах (тангенсы наклона или нормированные данные)
2. Расчет ε-трубки для параметра v
   $d_i = \varepsilon$ - в обычном случае
   $d_i = \varepsilon \cdot v_i$ - в адаптивном случае
3. Объединение векторов, которые попадают в одну ε-трубку

$$u \in K(v) \iff \forall i < \dim(v) \begin{cases} |u_i - v_i| < d_i \\ |u_{i+1} - v_{i+1}| < d_i \end{cases}$$

Программа Чубаровой Дарьи, Международный центр анализа и выбора решений

Алескеров Ф. Т. и др. Анализ паттернов в статике и динамике, часть 2: Примеры применения к анализу социально-экономических процессов //Бизнес-информатика. – 2013. – №. 4 (26). – С. 3-20.

Устойчивость сети

Что изменилось в сети с течением времени?

Ранжирование вершин

$$r_{ij}^t = \begin{cases} 1, & c_i^t - c_j^t > \varepsilon \\ 0, & otherwise. \end{cases}$$

- матрица интервального порядка

$$d\left(R^t, R^{t+1}\right) = \frac{\sum_{j \neq k}^{n} \left| r_{ij}^t - r_{ij}^{t+1} \right|}{n \cdot (n-1)}.$$

- расстояние Хемминга

Aleskerov F., Shvydun S. Stability and similarity in networks based on topology and nodes importance //Complex Networks and Their Applications VII: Volume 1 Proceedings The 7th International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2018 7. – Springer International Publishing, 2019. – С. 94-103.

Устойчивость сети

Что изменилось в сети с течением времени?

Топологическая структура

C - матрица влияния

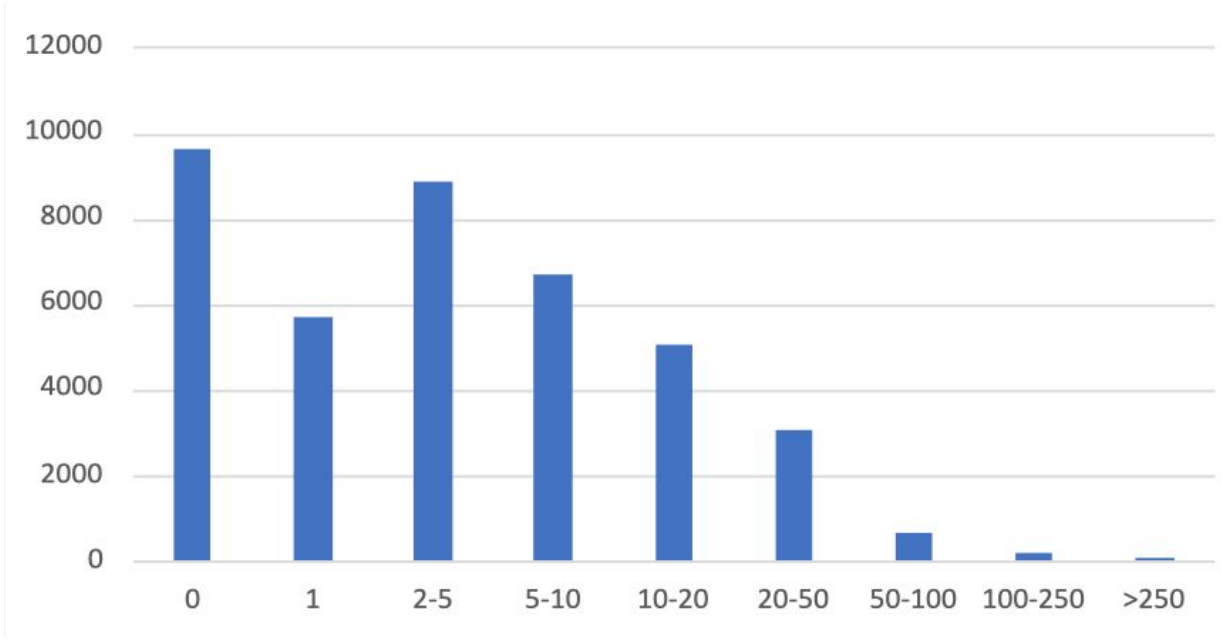
$$\delta\left(\widetilde{C}_t, \widetilde{C}_{t+1}\right) = \frac{\sum_{j,k}^{n}\left|\tilde{c}_{ij}^{t} - \tilde{c}_{ij}^{t+1}\right|}{n^2 \cdot \gamma}, \qquad \gamma = \max_{j,k}(\tilde{c}_{ij}^{t}, \tilde{c}_{ij}^{t+1}).$$

Стабильность

$$d\left(G^t, G^{t+1}\right) = \sqrt{\frac{d(R^t, R^{t+1})^2 + \delta\left(\widetilde{C}_t, \widetilde{C}_{t+1}\right)^2}{2}},$$

Aleskerov F., Shvydun S. Stability and similarity in networks based on topology and nodes importance //Complex Networks and Their Applications VII: Volume 1 Proceedings The 7th International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2018 7. – Springer International Publishing, 2019. – С. 94-103.

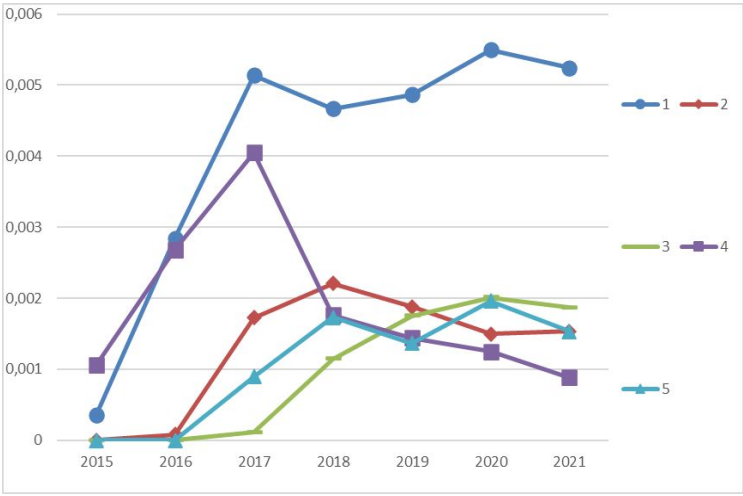
Результаты для статей



Распределение количества цитирований

Топ-5 статей по индексу In-degree

Название	DOI	In-degree	Ранг по Eigenvector	Ранг по PageRank
1. MDS clinical diagnostic criteria for Parkinson's disease	10.1002/MDS.26424	0.005	15	1
2. Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson’s Disease	10.1016/J.CELL.2016.11.018	0.0016	62	8
3. Epidemiology of Parkinson's disease	10.1007/S00702-017-1686-Y	0.00156	1224	5
4. MDS research criteria for prodromal Parkinson's disease	10.1002/MDS.26431	0.00152	13	3
5. The epidemiology of Parkinson's disease: risk factors and prevention	10.1016/S1474-4422(16)30230-7	0.0015	705	4



Изменение индекса по годам

Топ-5 статей по индексу TI, q = 0,1%

Название	DOI	In-degree	BI, q=0,1%	PI, q=0,1%	TI, q=0,1%
1. MDS clinical diagnostic criteria for Parkinson's disease	10.1002/MDS.26424	0,005	0,6307	0,9404	0,5254
2. Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson’s Disease	10.1016/J.CELL.2016.11.018	0,0016	0,0206	0,000096	0,0075
3. Epidemiology of Parkinson's disease.	10.1007/S00702-017-1686-Y	0,00156	0,0188	0,000093	0,0068
4. MDS research criteria for prodromal Parkinson's disease	10.1002/MDS.26431	0,00152	0,0175	0,000091	0,0064
5. The epidemiology of Parkinson's disease: risk factors and prevention	10.1016/S1474-4422(16)30230-7	0,0015	0,01704	0,00009	0,0062

Топ-5 статей по индексу TI, q = 1%

Название	DOI	In-degree	BI, q=1%	PI, q=1%	TI, q=1%
1. Vagotomy and subsequent risk of Parkinson's disease	10.1002/ANA.24448	0,0009	0,0204	0,068	0,0298
2. The clinical symptoms of Parkinson's disease	10.1111/JNC.13691	0,00089	0,0202	0,0673	0,0295
3. Oxidative stress and Parkinson’s disease	10.3389/FNANA.2015.00091	0,00087	0,0183	0,061	0,0267
4. Neuroinflammation in Parkinson’s disease and its potential as therapeutic target	10.1186/S40035-015-0042-0	0,00086	0,0177	0,0589	0,0258
5. Short chain fatty acids and gut microbiota differ between patients with Parkinson's disease and age-matched controls	10.1016/J.PARKRELDIS.2016.08.019	0,00082	0,0156	0,052	0,0228

Статьи с высоким значением индексов в 2021 году

Название	DOI	Год публикации	In-degree, ранг в 2015-2021	Индекс, ранг в 2021
1. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies	10.1016/S1474-4422(19)30320-5	2019	12	In-degree, 3
2. Diagnosis and Treatment of Parkinson Disease: A Review	10.1001/JAMA.2019.22360	2020	32	In-degree, 5
3. Extensive graft-derived dopaminergic innervation is maintained 24 years after transplantation in the degenerating parkinsonian brain	10.1073/PNAS.1605245113	2016	108	Eigenvector, 1
4. Personalized iPSC-Derived Dopamine Progenitor Cells for Parkinson's Disease	10.1056/NEJMOA1915872	2020	164	Eigenvector, 2
5. Pre-clinical study of induced pluripotent stem cell-derived dopaminergic progenitor cells for Parkinson's disease	10.1038/S41467-020-17165-W	2020	182	Eigenvector, 3
6. The future of stem cell therapies for Parkinson disease	10.1038/S41583-019-0257-7	2020	174	Eigenvector, 4

Сеть цитирований по аффилиациям

Строится по аффилиации первого автора

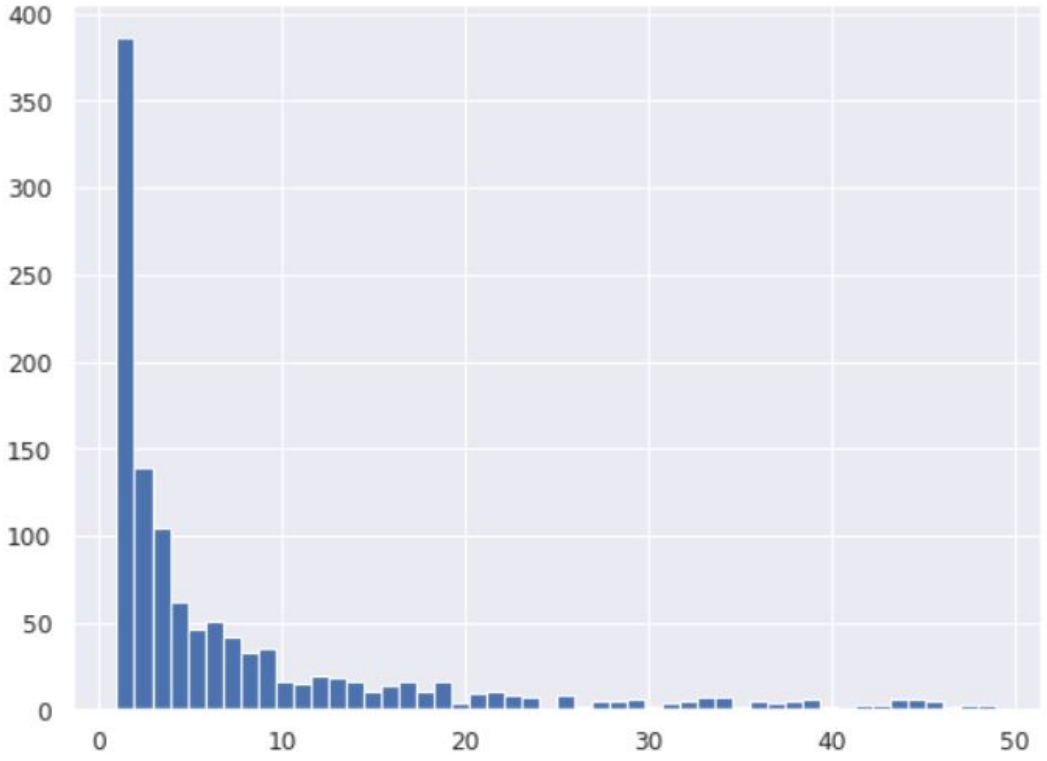
Количество вершин	3029
Количество ребер	128602
Количество компонент связности	5
Размер наибольшей компоненты связности	3025
Плотность графа	0.0244
Минимальное количество цитирований	0
Максимальное количество цитирований	3757
Среднее количество цитирований	79.6

Топ-10 аффилиаций по In-degree

Название	In-degree	Betweenness ранг	Eigenvector ранг	Pagerank ранг
1. National Institutes of Health	0,0156	5	1	1
2. UCL Institute of Neurology	0,0134	3	4	2
3. University of Cambridge	0,0131	1	3	4
4. University of Oxford	0,0122	2	2	3
5. University College London	0,0120	6	5	5
6. Northwestern University	0,0114	14	6	6
7. Harvard University	0,0100	8	8	7
8. University of Pennsylvania	0,0085	24	7	8
9. Katholieke Universiteit Leuven	0,0080	26	10	12
10. Karolinska Institutet	0,0079	10	9	9

Самоцитирования аффилиаций

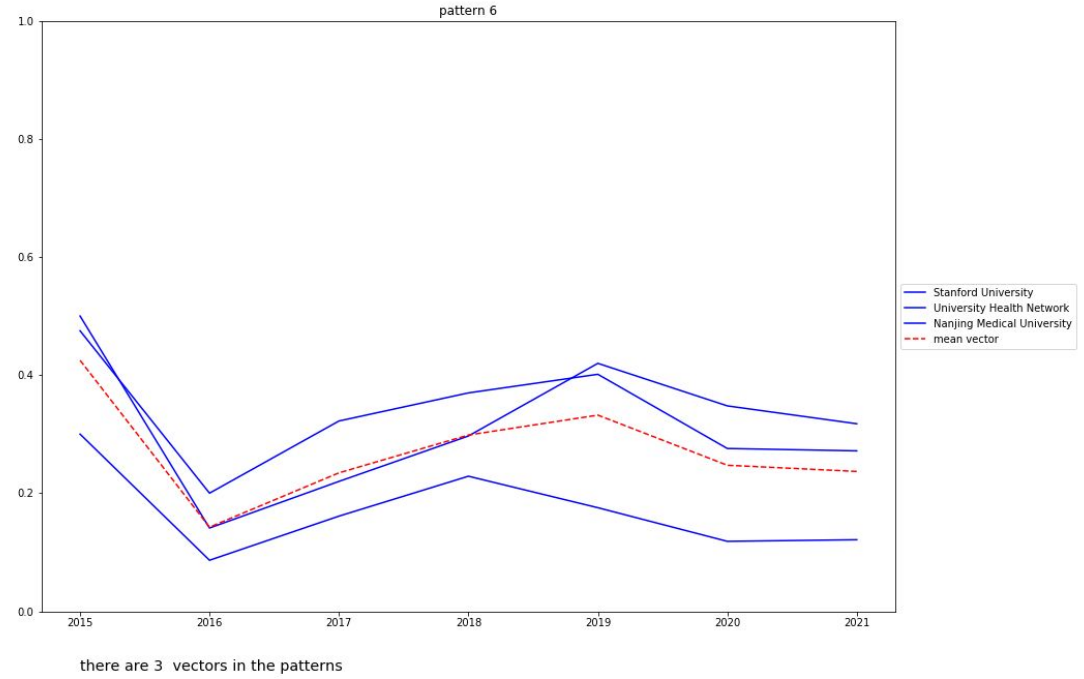
Распределение количества самоцитирований аффилиаций



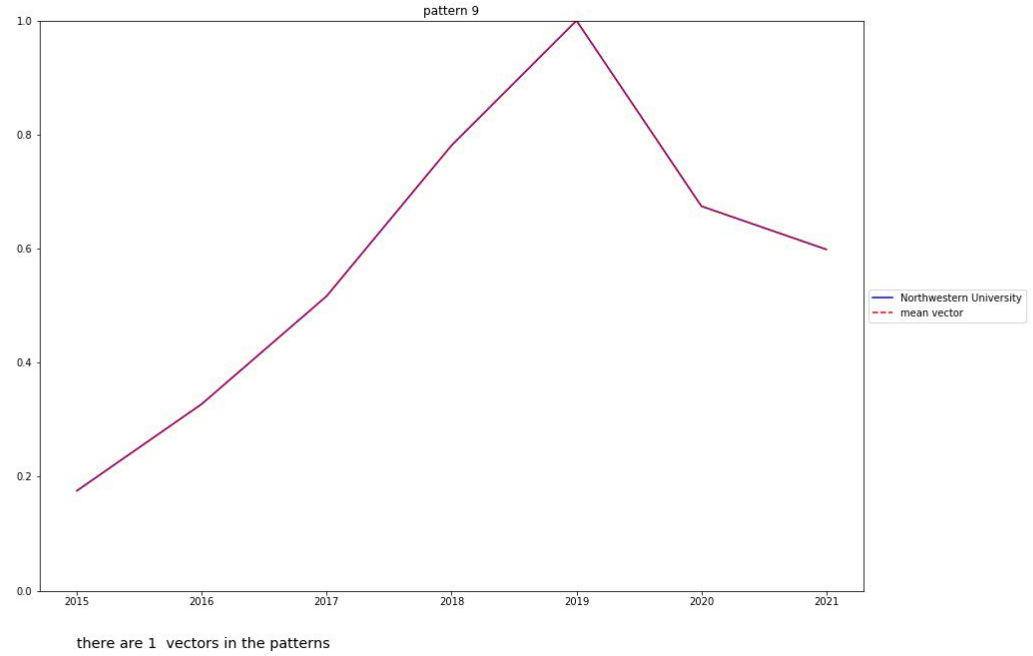
DAfN	In-degree	In-degree rank	weight	self_citations	self_citations_proportion
National Institutes of Health	0.015590	1.0	3757.0	419.0	0.111525
UCL Institute of Neurology	0.013416	2.0	3233.0	198.0	0.061243
University of Cambridge	0.013134	3.0	3165.0	396.0	0.125118
University of Oxford	0.012171	4.0	2933.0	359.0	0.122400
University College London	0.011959	5.0	2882.0	153.0	0.053088
Northwestern University	0.011449	6.0	2759.0	207.0	0.075027
Harvard University	0.010013	7.0	2413.0	149.0	0.061749
University of Pennsylvania	0.008490	8.0	2046.0	204.0	0.099707
Katholieke Universiteit Leuven	0.007996	9.0	1927.0	206.0	0.106902
Karolinska Institutet	0.007855	10.0	1893.0	178.0	0.094031

Паттерн-анализ аффилиаций

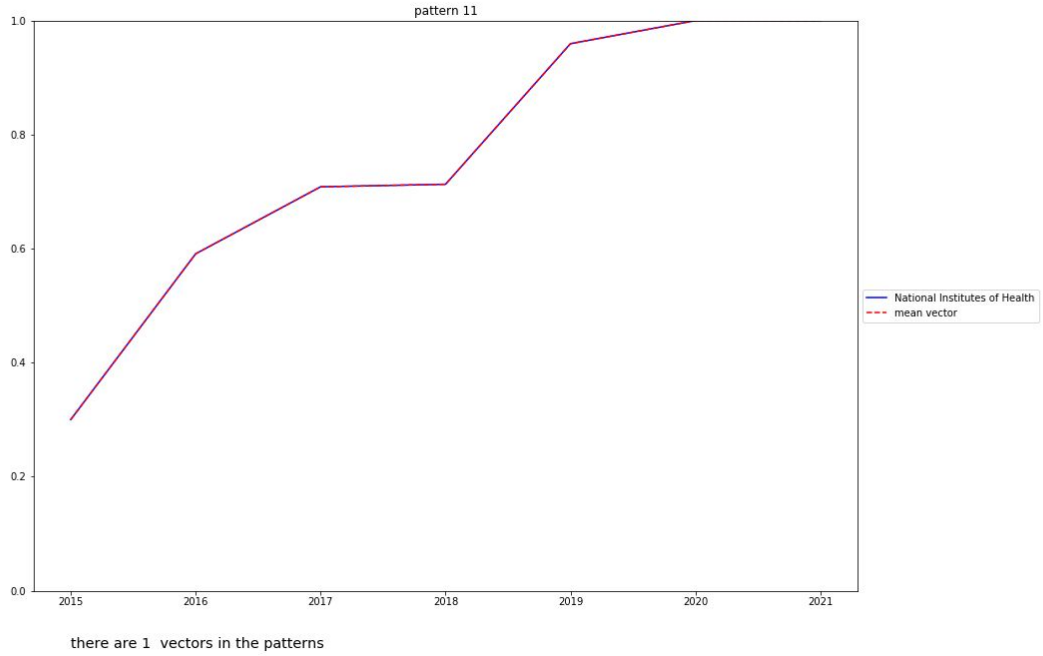
- 28 паттернов
- Более 90% аффилиаций в одном кластере
- 13 паттернов с одной организацией



Паттерн-анализ аффилиаций

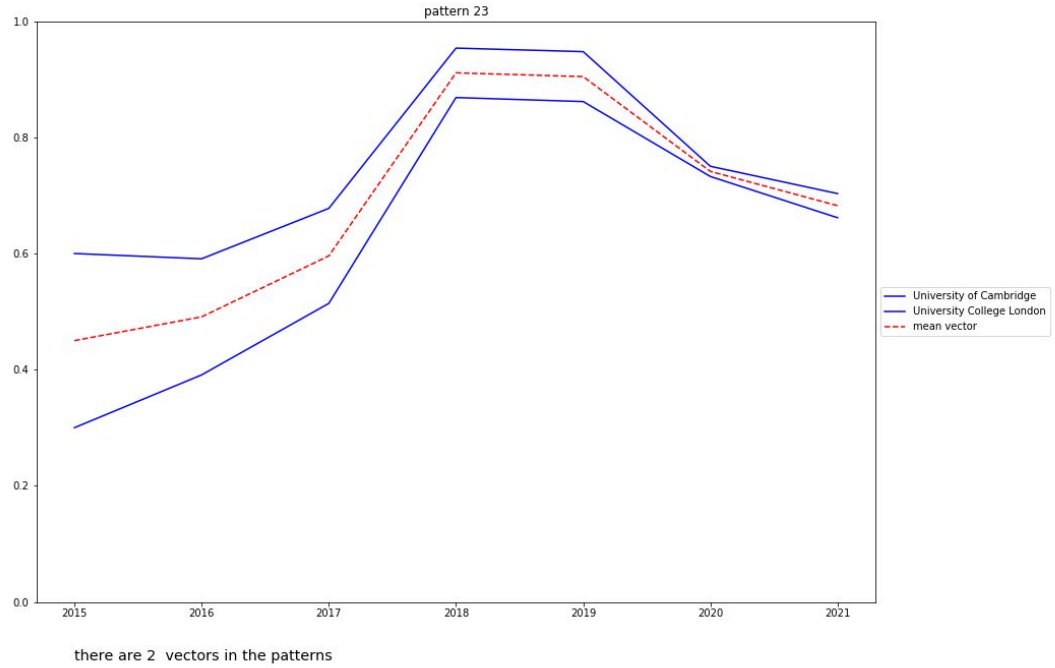


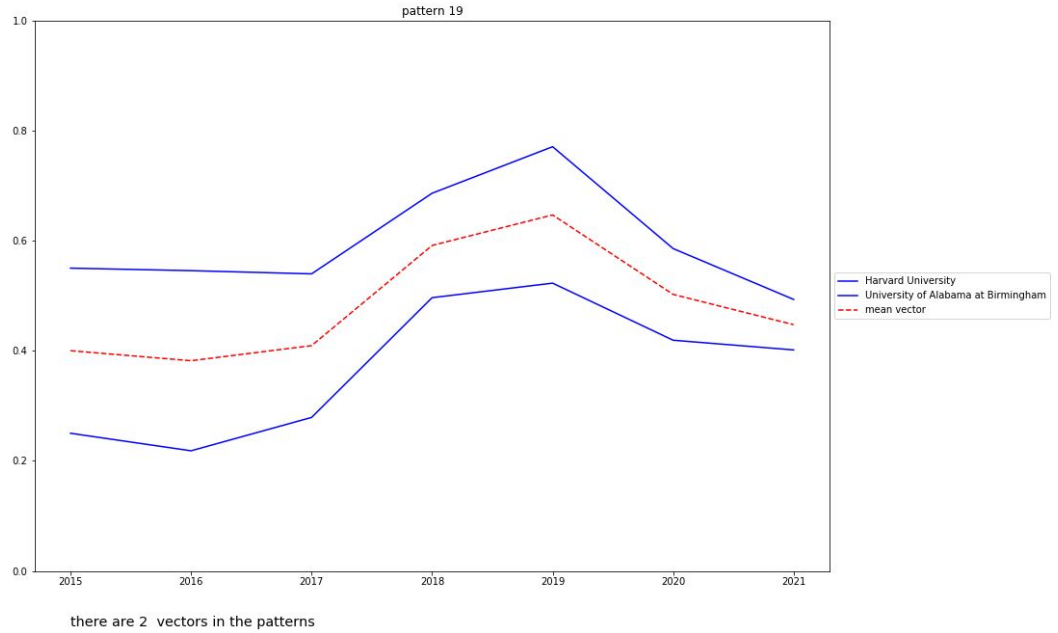
there are 1 vectors in the patterns



there are 1 vectors in the patterns

Паттерн-анализ аффилиаций





Заключение

- Собрано почти 40 тысяч публикаций и более чем 3 тысячи аффилиаций
- Подсчитаны классические индексы и недавно разработанные Bundle index и Pivotal index
- Работа представлена на конференции "The 12th International Conference on Network Analysis" в мае 2022 года
- Работа представлена на конференции "HCist 2022 - International Conference on Health and Social Care Information Systems and Technologies" в ноябре 2022 года

Результаты могут быть использованы исследователями БП для понимания ключевых областей исследований, их влияния на сообщество и для выявления новых направлений, которые раньше оставались без внимания.

Разработанная методика может применяться для выявления областей исследования для выгодных инвестиций.

Планы

1. Провести паттерн-анализ
2. Провести анализ аннотаций (семантическая близость публикаций)
3. Разработать модели по соавторам

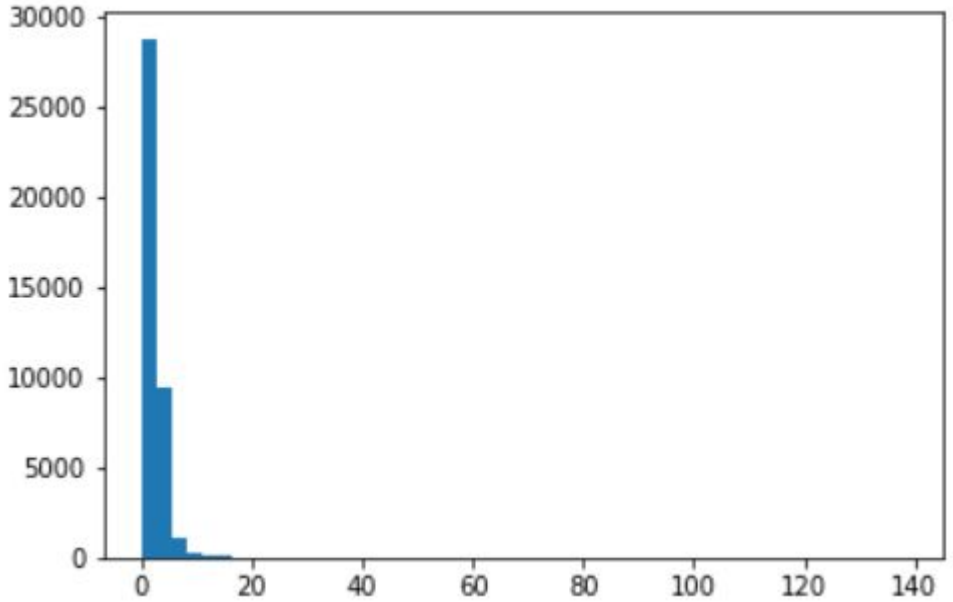
Сеть сотрудничества организаций

Вершины - организации

Ребра - написание организациями совместной статьи

Петли - авторы из одной организации написали статью

Распределение количества аффилиаций в одной статье



Термины

Тема	Количество терминов	Количество статей
Биохимия, генетика, клетка	75	25037
Лекарства	175	9600
Симптом	56	25139
Диагноз	27	10975
Методы исследования	19	9505
ЦНС	40	21600

Источники

1. Ruiz M. L., Benito-Le´on J. The top 50 most-cited articles in orthostatic tremor: A bibliometric review //Tremor and Other Hyperkinetic Movements. – 2019. – V. 9
2. Sorensen A. A., Weedon D. Productivity and impact of the top 100 cited Parkinson's disease investigators since 1985 //Journal of Parkinson's disease. – 2011. – V. 1. – No. 1. – С. 3-13
3.Ponce, F. A., Lozano, A. M. (2011). The most cited works in Parkinson's disease. Movement Disorders, 26(3), 380-390.
4. Kusumastuti S. et al. Successful ageing: A study of the literature using citation network analysis //Maturitas. – 2016. – Т. 93. – С. 4-12.
5. Martinez-Perez C. et al. Citation network analysis of the novel coronavirus disease 2019 (COVID-19) //International journal of environmental research and public health. – 2020. – Т. 17. – №. 20. – С. 7690.
6. Koseoglu M. A. Growth and structure of authorship and co-authorship network in the strategic management realm: Evidence from the Strategic Management Journal //BRQ Business Research Quarterly. – 2016. – Т. 19. – №. 3. – С. 153-170.
7. Aleskerov F., Khutorskaya O., Buldyaev A., Yamilov A. Parkinson's disease: Network analysis of publications' impact //2018 7th International Conference on Computers Communications and Control (ICCCC). – IEEE, 2018. – С. 82-85.
8. Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of mathematical sociology*, *2*(1), 113-120.
9. Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35-41.
10. Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, *30*(1-7), 107-117.
11. Aleskerov F., Yakuba V. (2020). Matrix-vector approach to construct generalized centrality indices in networks. // SSRN 3597948. Available at: https://ssrn.com/abstract=3597948
12. Алескеров Ф. Т. и др. Анализ паттернов в статике и динамике, часть 2: Примеры применения к анализу социально-экономических процессов //Бизнес-информатика. – 2013. – №. 4 (26). – С. 3-20.
13. Aleskerov F., Shvydun S. Stability and similarity in networks based on topology and nodes importance //Complex Networks and Their Applications VII: Volume 1 Proceedings The 7th International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2018 7. – Springer International Publishing, 2019. – С. 94-103.

Спасибо за внимание!