Использование QAP-анализа для определения зависимости между графами

В. А. Балаш, А. Р. Файзлиев, С. П. Сидоров, А. А. Гудков, А. Ж. Чекмарева, М. А. Левшунов

Саратовский госуниверситет

Саратов, 2-3 июля 2018 г.

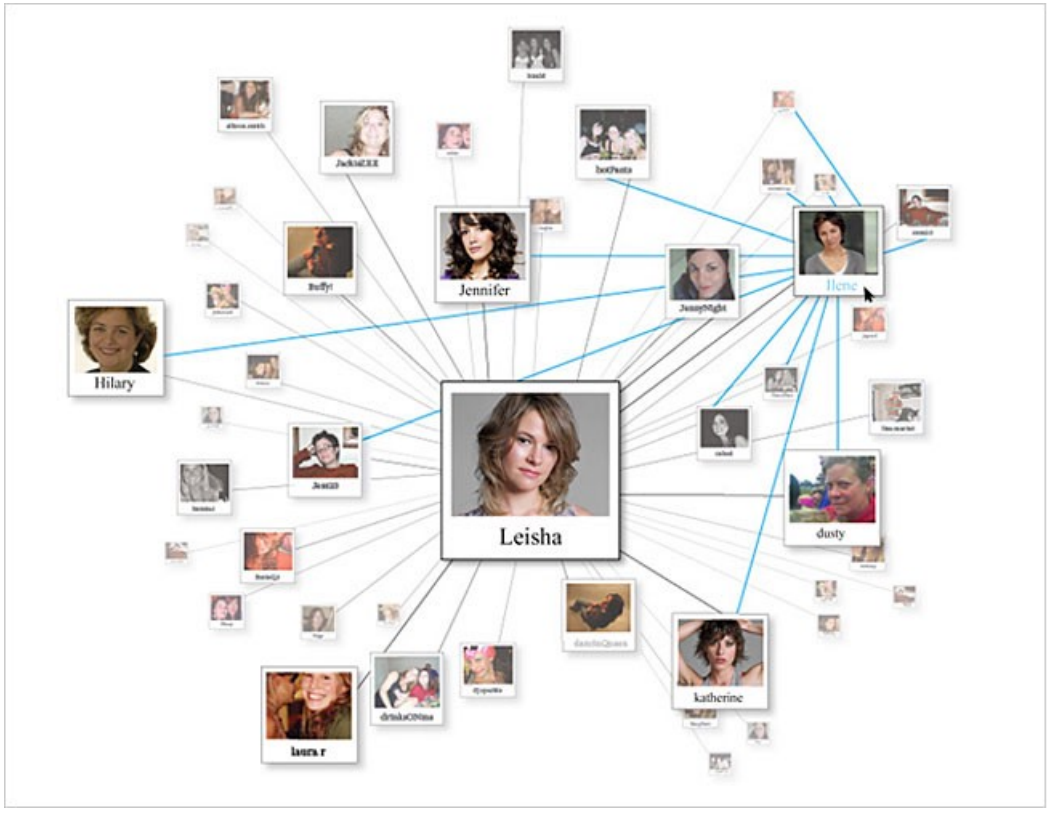
Работа поддержана РФФИ, грант 18-37-00060

Содержание

Цель работы

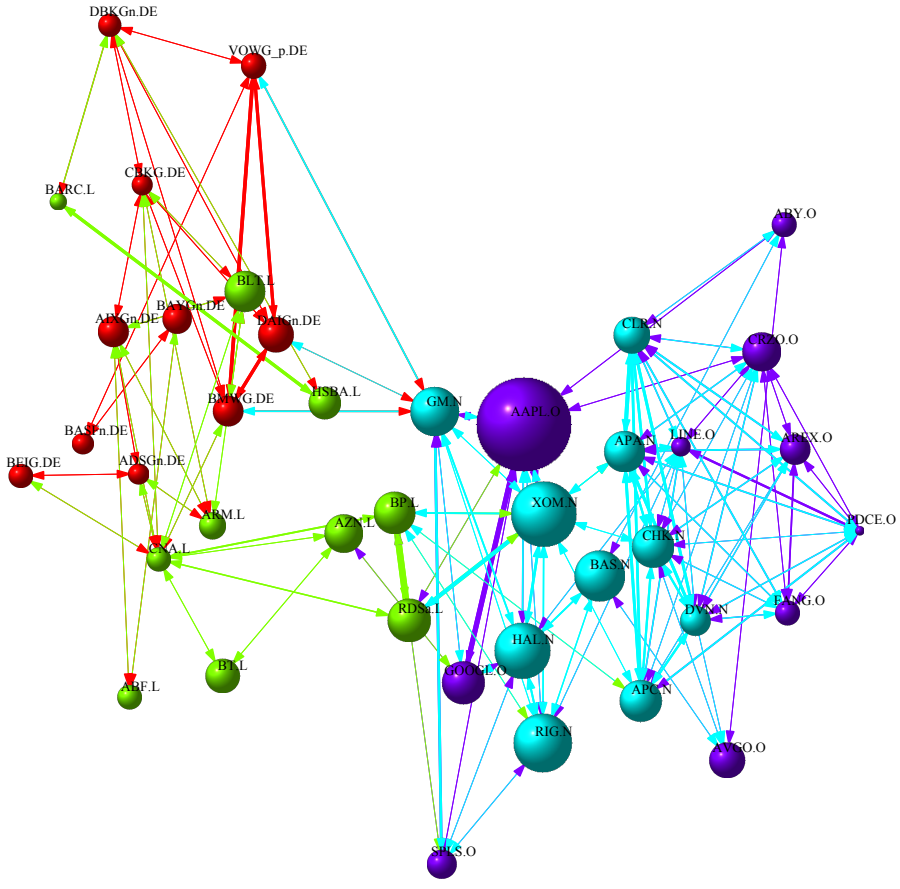
QAP анализ сети компаний на основе совместного упоминания их в новостных сообщениях.

Социальный граф (англ. Social graph) — это граф, узлы которого представлены социальными объектами, такими как пользовательские профили с различными атрибутами (например: имя, день рождения, родной город и т. д.), сообщества, медиа-контент и т. д., а ребра—социальными связями между ними.



Company Co-Mention Network

Сеть компаний, акции которых торгуются на крупных финансовых рынках, с точки зрения новостных информационных агентств.



Анализ со-упоминаний компаний

1. Сбор полных текстов всех экономических и финансовых новостей, опубликованных в течение некоторого периода времени.
2. Для каждого новостного сообщения собрать список компаний, которые упоминаются в данной новости.
3. Для всех наборов доступных новостей произвести подсчет ссылок для каждой пары со-упомянутых компаний.
4. Построить симметричную взвешенную матрицу со-упоминаний компаний.
5. Проанализировать матрицу совместного упоминания, и визуализировать и интерпретировать результаты.

Наиболее известные поставщики данных новостной аналитики

- RavenPack (http://www.ravenpack.com/).
- Media Sentiment (www.mediasentiment.com/).
- Thomson Reuters News Analytics (http://thomsonreuters.com).

TIMESTAMP_UTC	COMPANY	RELEVANCE	ESS	ENS	CSS	WLE	PCM	ECM	RCM	VCM	NIP
01.01.2005 14:00	US/BXS	100	49	100	52	50	50	100	50	50	36
02.01.2005 3:37	HK/2388	100	54	100	50	50	50	50	50	50	44
02.01.2005 13:55	GB/OOM	100	66	100	50	50	50	50	50	50	23
02.01.2005 15:15	CH/ROG	100	76	100	52	50	50	100	50	50	36
02.01.2005 15:15	DE/BAY	100	49	100	52	50	50	100	50	50	36
02.01.2005 15:30	CH/ROG	100	76	75	52	50	50	100	50	50	36
02.01.2005 15:30	DE/BAY	100	49	75	52	50	50	100	50	50	36
03.01.2005 0:58	KR/005380	100	50	100	50	50	50	50	50	50	52
03.01.2005 0:59	KR/005380	100	50	75	50	50	50	50	50	50	51
03.01.2005 1:00	TW/2330	100	50	100	53	100	50	100	50	50	57
03.01.2005 1:00	KR/005380	100	50	56	50	50	50	50	50	50	61
03.01.2005 1:18	KR/005380	100	50	42	50	50	50	50	50	50	52
03.01.2005 1:33	KR/005380	100	50	32	50	50	50	50	50	50	53
03.01.2005 1:45	CN/000898	100	63	100	50	50	50	50	50	50	41
03.01.2005 1:53	HK/0022	100	69	100	50	50	50	50	50	50	76

Table 1. News / companies

	c1	c2	c3	c4	c5	c6	c7
N1	+	+	+				
N2			+	+			
N3				+	+	+	
N4						+	+
N5	+	+					
N6				+		+	
N7					+	+	+
N8	+	+					

Table 2. The matrix of weights

	c1	c2	c3	c4	c5	c6	c7
c1	0	3	1	0	0	0	0
c2	3	0	1	0	0	0	0
c3	1	1	0	1	0	0	0
c4	0	0	1	0	2	3	0
c5	0	0	0	2	0	2	0
c6	0	0	0	3	2	0	1
c7	0	0	0	0	0	1	0

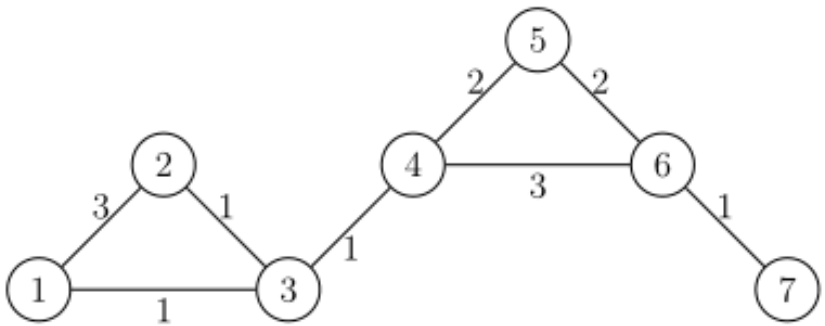


Fig. 1. Company co-mention network

Описательная статистика данных новостной аналитики с 1 февраля 2015 года по 28 февраля 2015 года (20 торговых дней)

δ, day	1
n	28
Sum	234736
Mean	8383.43
Minimum	262
Maximum	15069
St. deviation	5415,21
Median	10653
Skewness	-0.60
Kurtosis	1.75

Показатели анализа сети

1. Frequency - общее количество новостей в которых со-упоминается компания.
2. Degree - число связей компании.
3. Degree Centrality - определяется как количество связей компании, деленное на максимально возможное количество связей
4. Closeness Centrality - позволяет определить насколько близко узел относится ко всем другим узлам.
5. Betweenness Centrality - это доля путей, содержащих данный узел, который соединяет пару узлов в сети.
6. Eigenvector Centrality - подсчитывает общее количество узлов, которые примыкают к данному узлу, учитывая общее количество смежных узлов и важность каждого из смежных узлов. В некотором смысле, связи с влиятельными людьми дадут больше, чем связь с менее важными людьми

Сектора экономики

Потребительские товары длит. польз.
Потребительские товары
Промышленные предприятия
$\longrightarrow$ Продукты (2007 компаний)

Энергетика
Сырье
$\longrightarrow$ Ресурсы (1398 компаний)

Финансы
Здравоохранение
Телекоммуникационные услуги
Коммунальные услуги
$\longrightarrow$ Сектор услуг (2375 компаний)

Информационные технологии $\longrightarrow$ ИТ сектор (548 компаний)

Таблица: Компании с высокой частотой для четырех секторов экономики

Компании	Frequency	Degree	Degree Centrality $\times 10^1$	Betweenness Centrality $\times 10^1$	Eigenvector Centrality
Продукты					
General Motors Co	1051	143	0.71	0.72	1.000
Ford Motor Co	691	90	0.45	0.26	0.877
Volkswagen AG	668	101	0.50	0.23	0.572
Ресурсы					
Apache Corp	1583	118	0.84	0.04	0.966
Continental Resources Inc	1538	101	0.72	0.06	1.000
Pioneer Natural Resources Co	1422	70	0.50	0.00	0.984
Сектор услуг					
JPMorgan Chase & Co	882	137	0.58	0.43	1.000
Citigroup Inc	837	128	0.54	0.43	0.889
Bank of America Corp	757	142	0.60	0.69	0.763
ИТ сектор					
Apple Inc	805	127	2.32	2.59	1.000
Alphabet Inc	472	67	1.22	0.79	0.927
Twitter Inc	371	54	0.99	0.27	0.640

Таблица: Компании с высокой частотой для трех бирж

Компании	Frequency	Degree	Degree Centrality $\times 10^1$	Betweenness Centrality $\times 10^2$	Eigenvector Centrality
London SE					
Barclays PLC	484	42	1.72	3.37	0.374
Aviva PLC	429	47	1.93	5.20	0.308
BHP Billiton PLC	363	81	3.33	4.29	1.000
BP PLC	313	53	2.18	3.43	0.536
Associated British Foods PLC	303	68	2.79	3.00	0.733
New-York SE					
Continental Resources Inc	1594	178	2.08	0.50	0.720
Apache Corp	1588	184	2.15	1.02	0.733
Anadarko Petroleum Corp	1386	217	2.53	1.96	0.831
Basic Energy Services Inc	1336	252	2.94	1.05	1.000
Halliburton Co	1281	247	2.88	1.52	0.924
Tokyo SE					
Bombardier Inc	515	154	5.26	6.32	0.994
Calfrac Well Services Ltd	489	105	3.59	0.59	0.710
Alara Resources Ltd	464	120	4.10	1.08	0.883
Newalta Corp	428	147	5.03	1.86	1.000
Strad Energy Services Ltd	414	101	3.45	0.68	0.633

Корреляционный анализ QAP (Krackhardt,1988)

Цель

Проверить гипотезу о воспроизводстве в сетевом графе со-упоминаний компаний их отраслевой и биржевой принадлежности.

В дополнение к матрице со-упоминания были созданы 3 матрицы смежности:
1) Описывающие связь компаний через сектора экономики, к которым принадлежат компании (10 и 4 отрасли экономики).
2) Описывающую связь компаний через их биржевую принадлежность.

Стандартный корреляционный анализ не подходит для таких данных (не являются независимым друг от друга). Это противоречит одной из основных допущений линейного регрессионного анализа.

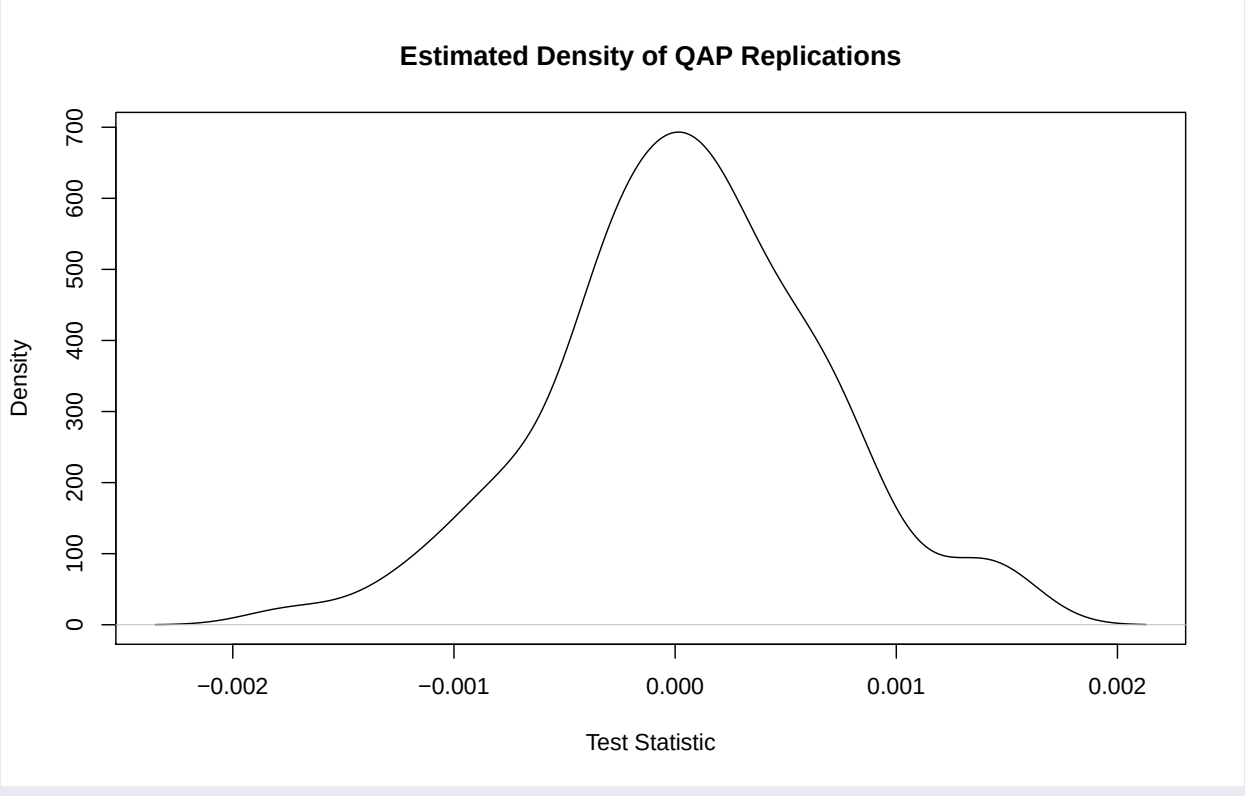
QAP(quadratic assignment procedure)

Данные матрицы сравниваются с вычислением коэффициента корреляции Пирсона. Далее данный процесс повторяется, переставляя случайным образом столбцы и строки, чтобы найти корреляцию.

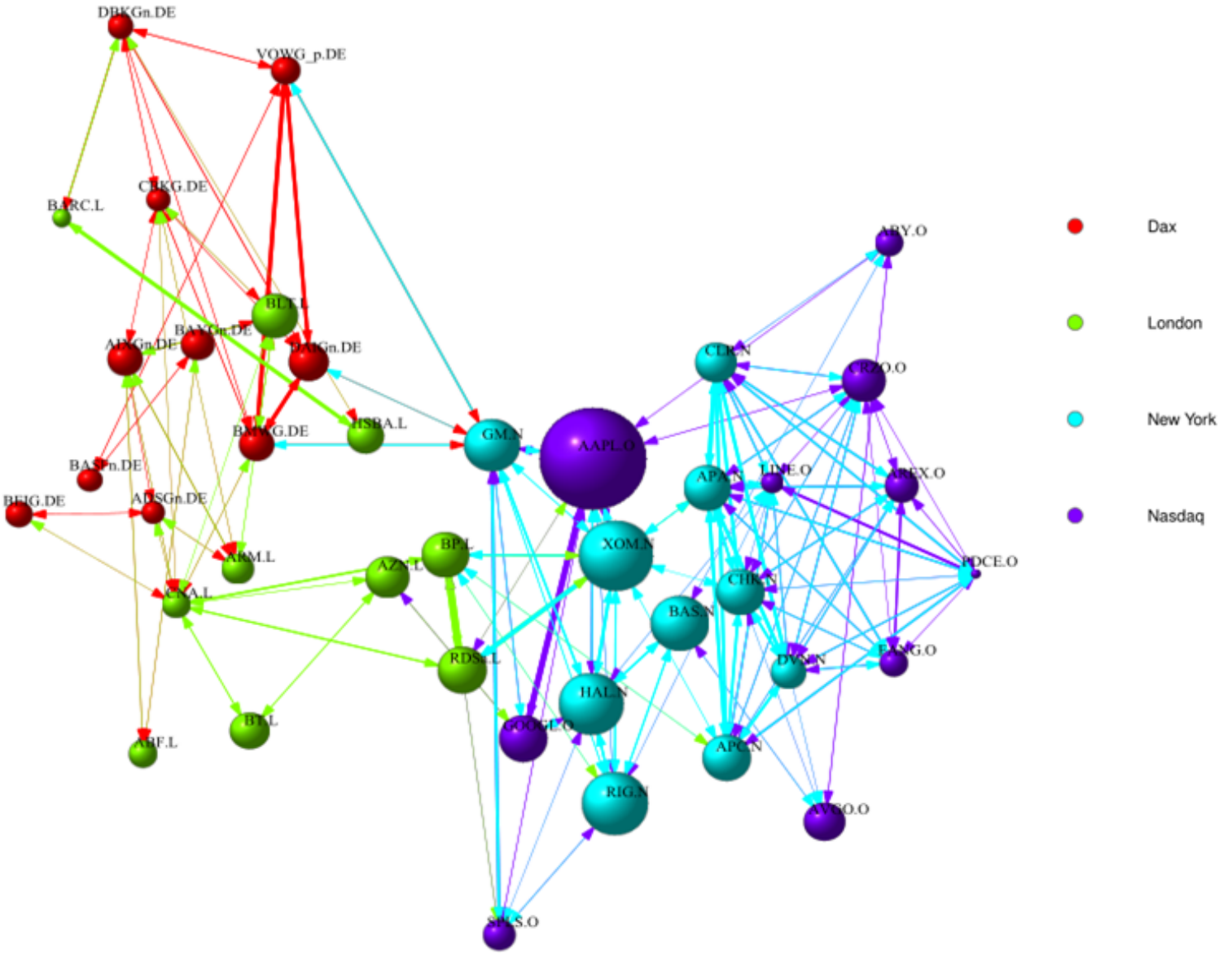
QAP (quadratic assignment procedure)

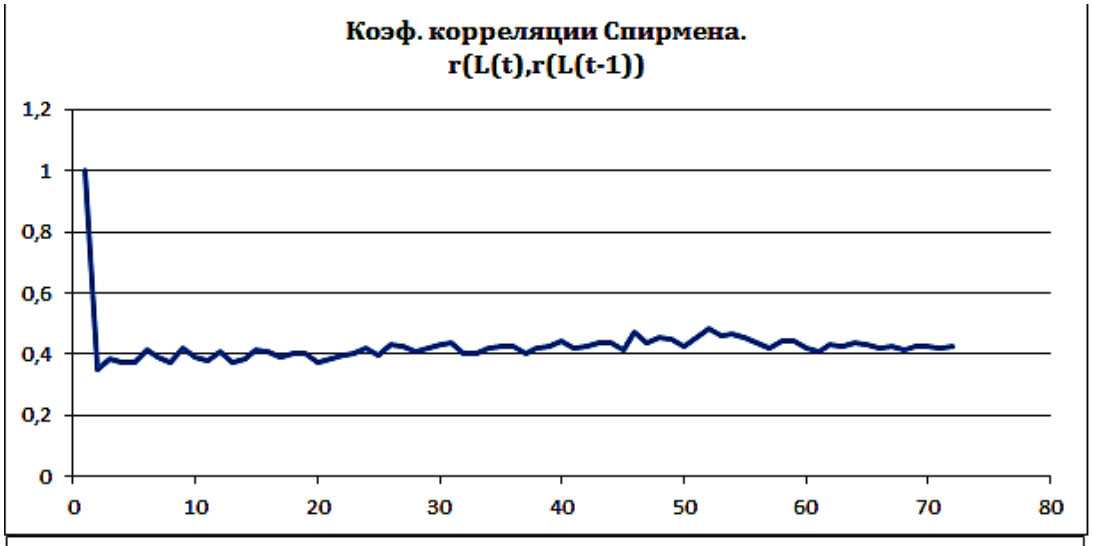
1) Между сетью со-упоминаний и сетью связей компаний по биржам:
r = 0,053, p = 0,000.
2) Между сетью со-упоминаний и сетью связей по секторам (для 10 секторов):
r = 0,020, p = 0,000.

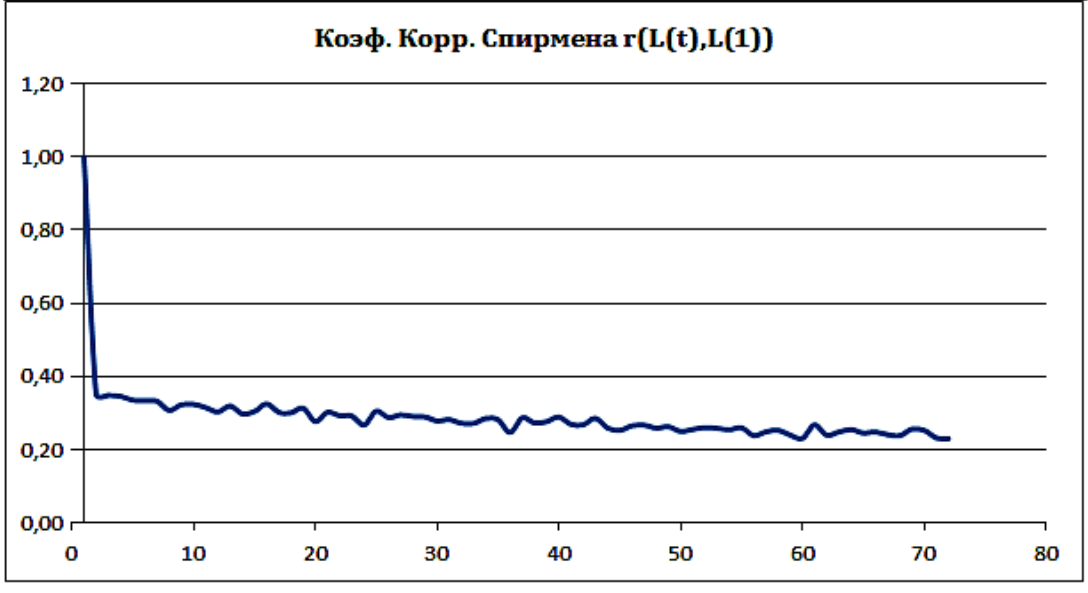
Оценочная плотность повторов QAP для сети ассоциаций по биржам



Фрагмент графа со-упоминаний




Коэф. корреляции Спирмена.
r(L(t),r(L(t-1))
1,2
1
0,8
0,6
0,4
0,2
0
0
10
20
30
40
50
60
70
80

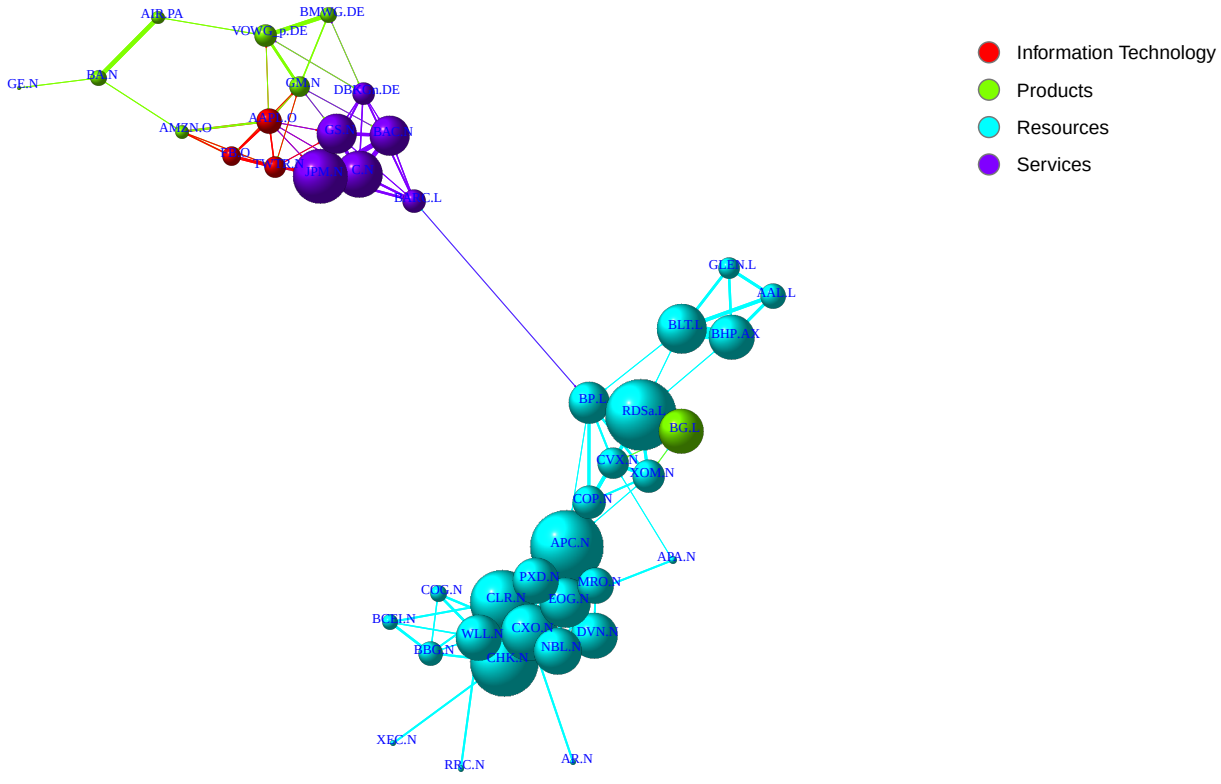

Коэф. Корр. Спирмена r(L(t),L(1))
1,20
1,00
0,80
0,60
0,40
0,20
0,00
0
10
20
30
40
50
60
70
80

$$L = T + S + \varepsilon$$

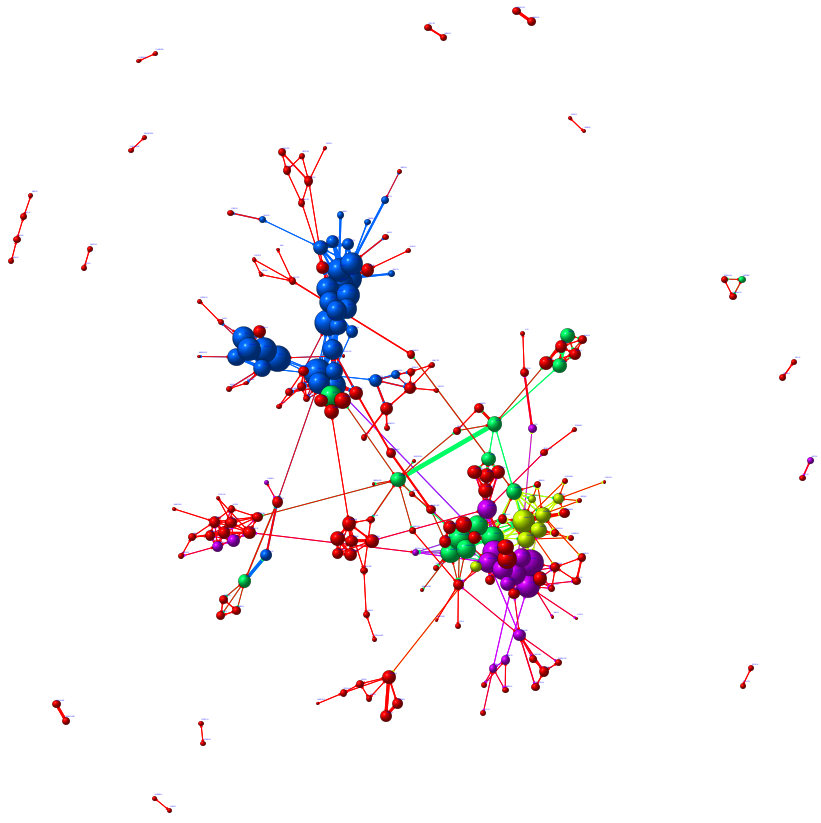
$$L_t = T + Event_t + \beta_1 L_{t-1} + u_t, u_t \sim iid(0, \sigma^2)$$

1. $P(L_{ij}^t > 0) > \gamma, \gamma = 0.9, 0.8, \ldots$ - устойчивая связь.
2. $P(L_{ij}^t > 0) < \gamma', \gamma' = 0.1, 0.2, \ldots$ - не устойчивая.

Пример стабильной части графа (50 компаний)



Пример стабильной части графа (250 компаний)



Благодарю за внимание!