Методология измерения научной эффективности научно-технического центра в нефтегазовой отрасли.

Семинар лаборатории МУСС ФКН ВШЭ.

Федор Краснов, к.т.н.

Газпромнефть НТЦ

2019

План презентации

Общая характеристика исследования

- Почему научно-технические центры в нефтегазовой отрасли?
- Почему научная эффективность?

Постановка эксперимента для прямой и обратной задач.

Методы исследования.

- Анализ социальных сетей.
- Анализ естественного языка

Научная новизна

Результаты эксперимента

Почему научно-технические центры в нефтегазовой отрасли?

1. Качественный скачок в структуре и динамике развития производительных сил обеспечивается деятельностью отраслевых научно-технических центров (НТЦ).
2. Количество НТЦ в энергетической отрасли растёт из года в год, а по мере исчерпания запасов легко добываемой нефти роль научной составляющей в ее добыче возрастает.
3. Эффективность деятельности НТЦ является ключевой характеристикой, нуждающейся в оценке и планировании.
4. Современные НТЦ представляют собой научно-проектную структуру, которая полностью интегрирована в производство. Оценка деятельности таких НТЦ нуждается в пересмотре.

Газпром нефть

Почему научная эффективность?

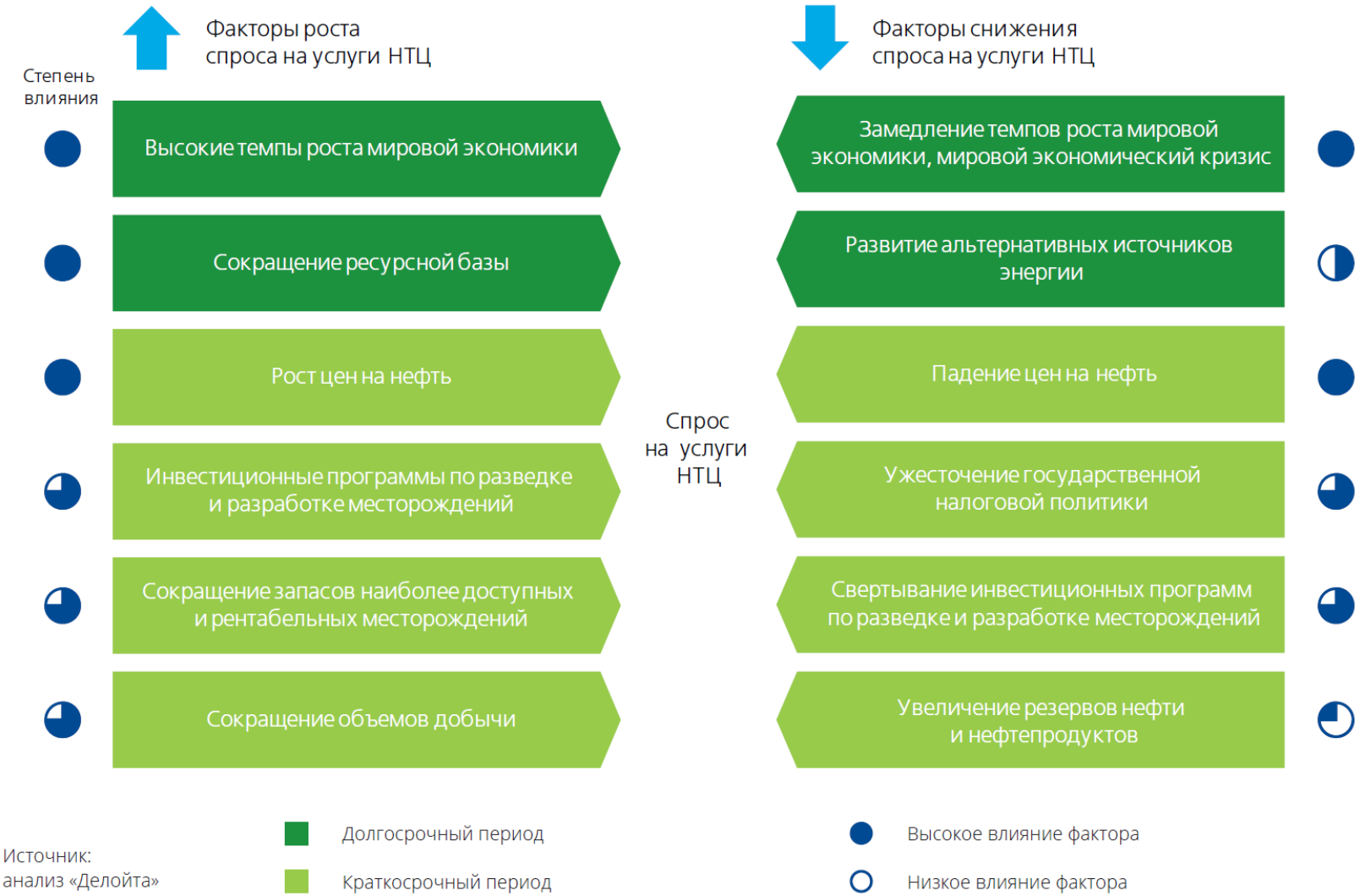


Рис. 1: Спрос на услуги НТЦ.

Основной результат

Создана аппаратно-аналитическая платформа СПО “НауБот” для моделирования научной деятельности НТЦ на основе алгоритмов машинного обучения и иммитационного моделирования.

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2017661666

СПО «НауБот»

Правообладатель: ***Общество с ограниченной ответственностью «Газпромнефть Научно-Технический Центр» (ООО «Газпромнефть НТЦ») (RU)***

Автор: ***Краснов Федор Владимирович (RU)***

Заявка № **2017618802**
Дата поступления **31 августа 2017 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***18 октября 2017 г.***

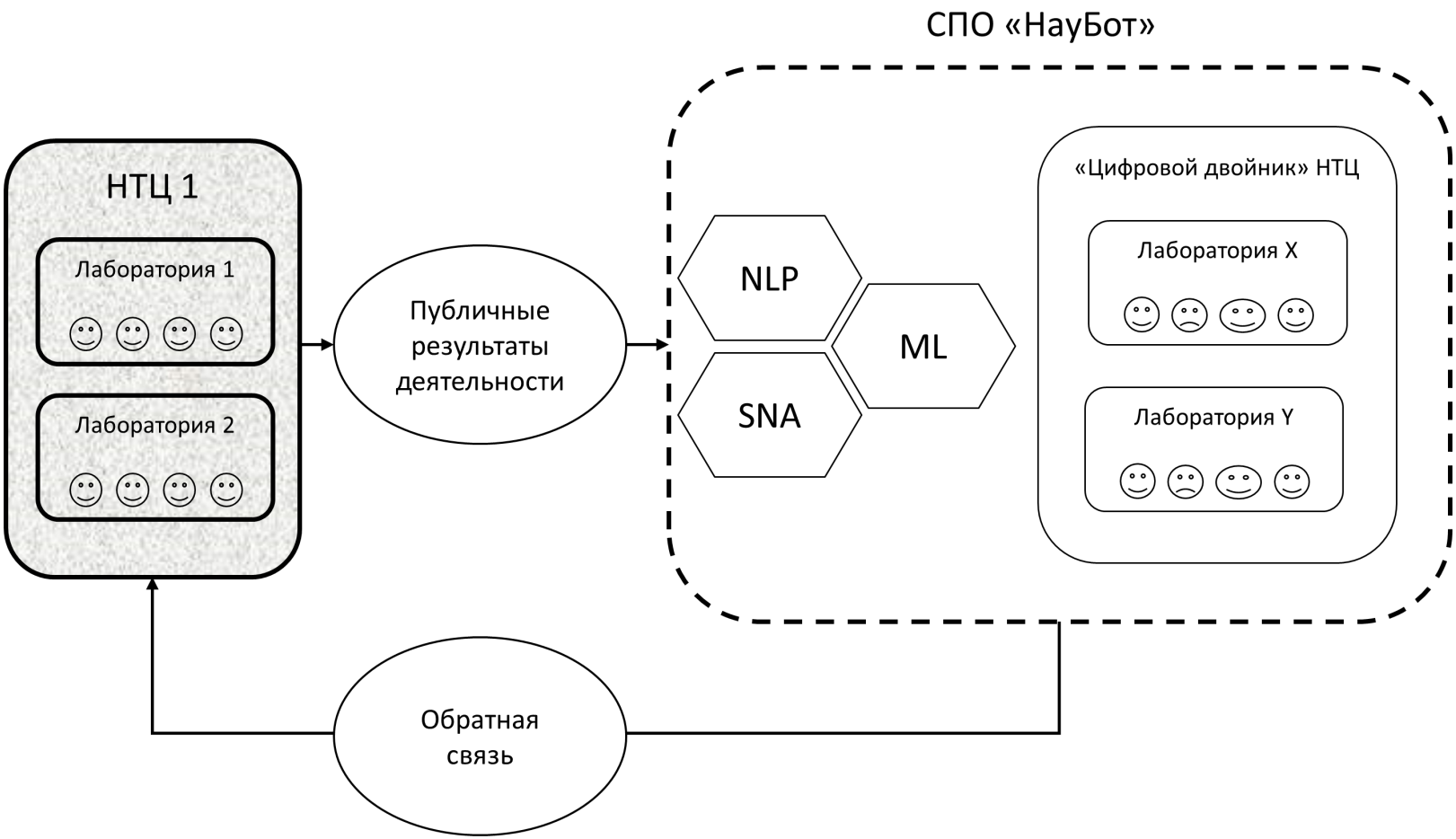
Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

Результаты работы

1. Предложена формализация процесса самоорганизации команд для достижения определённой цели;
2. Разработан детальный алгоритм образования соавторств;
3. Исследована временная зависимость структуры соавторств;
4. Создана модель для прогнозирования соавторств;
5. Создана модель научных направления развития НТЦ на основе публичных данных о публикационной активности сотрудников;
6. Создана модель движения персонала в организации и модель выполнения наукоемких заданий;
7. Разработан математический аппарат построения графов соавторства на основе двудольного графа;
8. Построен “цифровой двойник” НТЦ.

Архитектура СПО “НауБот”



Постановка эксперимента для прямой и обратной задач.

- *Изучение деятельности НТЦ по внешним проявлениям.* К внешним проявлениям относятся цифровые артефакты деятельности организации - это опубликованные научные статьи, материалы конференций, информационные сайты в сети Интернет и новости о компании.
- *Изучение НТЦ изнутри.* К исследованиям в этом направлении относятся моделирование научной деятельности, эффективность производственных процессов, самоорганизации малых творческих коллективов и модели персонала научной организации.

Социальная сеть: Граф соавторов (I).

Социальная сеть: Метрики графа (II).

- Для ребер
  - Common Neighbours (CN)
  - Salton Index (SI)
  - Jaccard Index (JI)
  - Hub Promoted Index (HPI)
  - Hub Depressed Index (HDI)
  - Leicht-Holme-Newman Index (LHN1)
  - Preferential Attachment Index (PA)
  - Adamic-Adar Index (AA)
  - Resource Allocation Index (RA)
- Для вершин
  - Degree centrality
  - Betweenness centrality
  - Closeness centrality
  - Harmonic centrality
  - Clustering

NLP

Допустим у нас есть вероятность последовательности из n слов $P(w_1, \ldots, w_n)$, такая, что вероятность третьего слова $P(w_3)$ равна $P(w_3|w_1, w_2)$. Тогда следующее выражение определяет вероятностную модель текста.

$$P(w) = P(w_1, w_2, \ldots, w_n) = \prod_i^n P(w_i|w_1, w_2, \ldots, w_{i-1}) \qquad (1)$$

Так как вычисление $P(w)$ представляет сложность O^n, то современные исследования текста используют представление $P(w)$, как однородной Цепи Маркова и строят приближенные модели:

- Униграмная модель $P(w_1, w_2, \ldots, w_n) \approx \prod_i P(w_i)$
- Биграмная модель $P(w_i|w_1, w_2, \ldots, w_{i-1}) \approx \prod_i P(w_i|w_{i-1})$

Теория суррогатного моделирования: этапы создания моделей.

- Характеристика объекта $Z = \Phi(X, Y)$, где переменная X описывает сам объект, а переменная Y задает условия функционирования.
- Функция Φ является неизвестной, и для ее вычисления проводятся вычислительные эксперименты.
- Измерения $\Xi = \{X_i, Y_i, Z_i = \Phi_i(X_i, Y_i), i \in \mathbb{R}\}$, где значение $Z_i = \Phi_i(X_i, Y_i)$ характеристики Z получено методом M_i для объекта, имеющего описания X_i, в условиях функционирования Y_i.
- По известному множеству Ξ с помощью строится функция $\Phi^s(X, Y)$, значение которой принимаются в качестве приближенного значения характеристики Z.

Если все значения в множестве Ξ получены при помощи одной и той же модели M и $\Phi^s(X, Y) \approx \Phi^m(X, Y)$, то построенная функция Φ^s может рассматриваться как “заменитель” (суррогат) функции Φ^m.

Байесовские методы для определения параметров НТЦ

Пусть дана функция $\Phi(x)$ и нам нужно найти x при котором она достигает максимума $\Phi(x) \longrightarrow \max_x$. Добавим условие при котором расчет каждого значения $\Phi(x)$ это ресурсоемкая задача. Такое условие встречается в следующих случаях:

- x - это географические координаты скважины, а $\Phi(x)$ - это количество нефти, которое можно добыть, пробурив скважину с координатами x. В таком случае одно значение $\Phi(x)$ стоит миллионы рублей;
- x - это гиперпараметры искусственной нейронной сети глубокого обучения, $\Phi(x)$ - это целевая метрика точности предсказания. В этом случае одно значение $\Phi(x)$ будет занимать месяцы работы;

EM-алгоритм с регуляризацией (I)

Байесовская модель для постериорного распределения скрытых тематик в тексте может быть записана в следующем виде.

$$p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d)\, p(w_{dn}|z_{dn})$$

$$p(\theta_d) \sim Dir(\alpha)$$

$$p(z_{dn}|\theta_d) = \theta_{dz_{dn}}$$

$$p(w_{dn}|z_{dn}) = \Phi_{z_{dn}w_{dn}}$$

$$\sum_{w} \Phi_{tw} = 1$$

$$\Phi_{tw} \geqslant 0$$

Таким образом, что W - это текстовые данные , Φ - распределение слов в каждой тематике, Z - распределение тематик для каждого слова, Θ - распределение тематик в документе.

EM-алгоритм с регуляризацией (II)

Оптимизационная задача для поиска скрытых тематик выглядит следующим образом:

$$P(W|\Phi) \to max_{\Phi} \tag{2}$$

Для использования EM-алгоритма выпишем явно уравнения для E-шага и M-шага:

E-шаг:

$$\mathcal{KL}(q(\Theta)\, q(Z)||p\,(\Theta, Z|W)) + \mu * R(q(\Theta)\, q(Z)) \to \underset{q(\Theta)\, q(Z)}{\text{minimize}} \tag{3}$$

M-шаг:

$$\mathbb{E}_{q(\Theta)\, q(Z)} \log p\,(\Theta, Z, W) \to \underset{\Phi}{\text{maximize}} \tag{4}$$

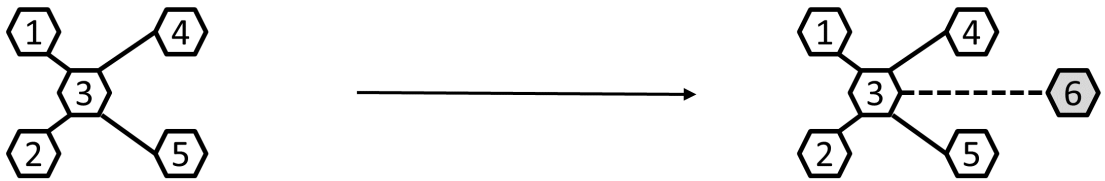
Имитационное моделирование

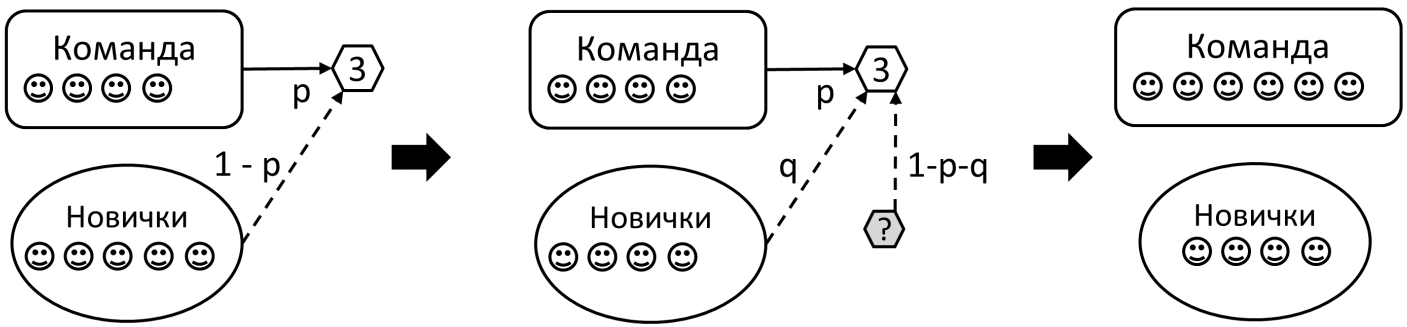
Допустим, что в отраслевой научно-исследовательской организации Ω работают лаборатории λ_i , где $i \in (1 \dots N_\lambda)$. Обозначим множество лабораторий $\Lambda = \{\lambda_i, \dots, \lambda_{N_\Lambda}\}$. В лабораториях работают научные сотрудники $A = \{a_i, \dots, a_{N_A}\}$. Обозначим множество тематик t_i, где $i \in (1, \dots, N_T)$, по которым организация Ω ведет НИР как $T = \{t_1, \dots, t_{N_T}\}$. Тогда деятельность организации Ω по выполнению НИР может быть описана следующими компонентами :

$$\mathbb{M}_\Omega = \Big\{ S, \Xi, \Psi, E \Big\}, \text{ где } S = \{\Lambda, A, T, P, X\} \tag{5}$$

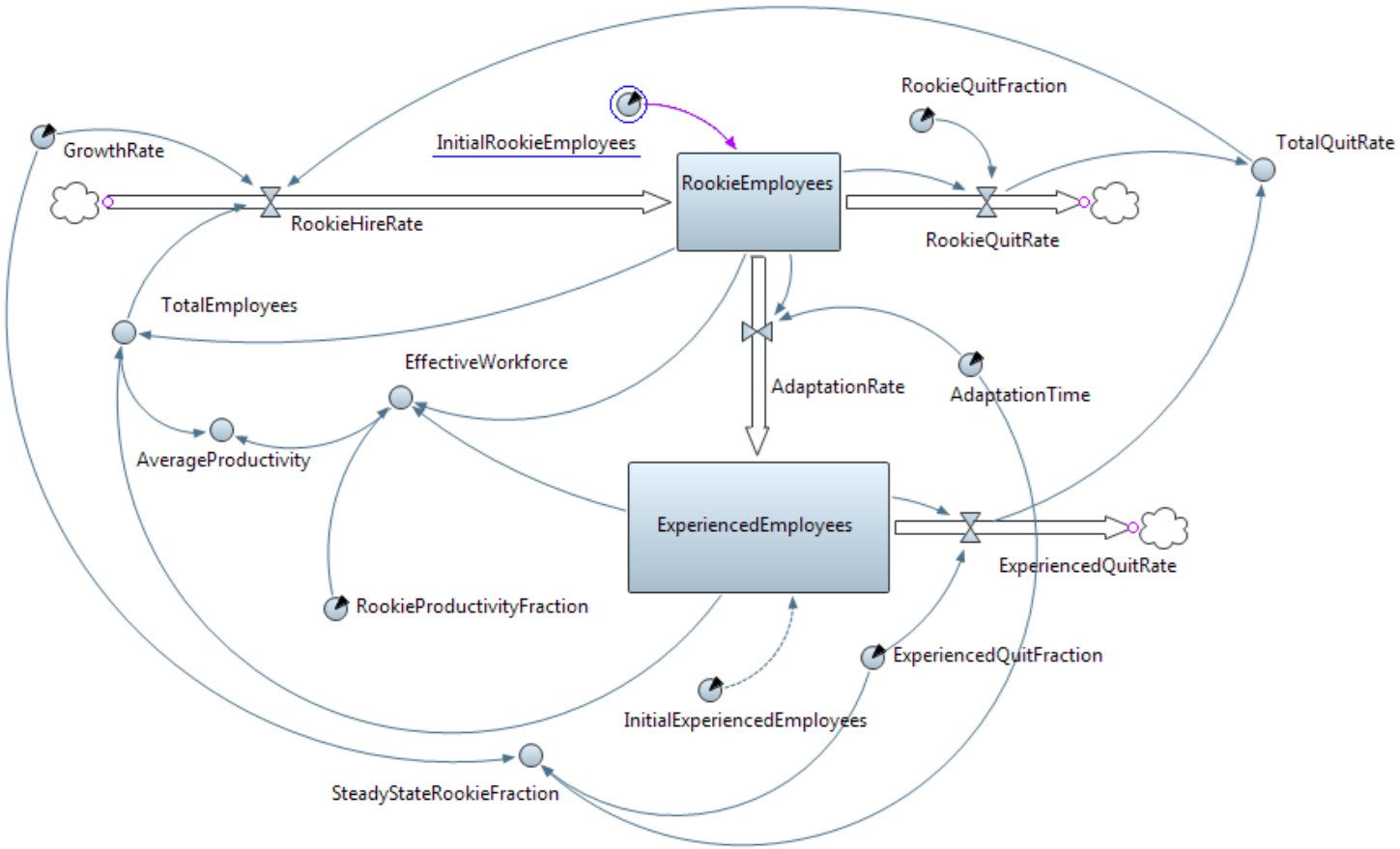
- $\Xi = \{\xi_1, \dots, \xi_{N_\Xi}\}$ – множество связей между субъектами,
- $\Psi = \{\psi_1, \dots, \psi_{N_\Psi}\}$ – множество действий субъектов,
- $P = \{\rho_1, \dots, \rho_{N_P}\}$ – множество научных работ,
- $X = \{\chi_1, \dots, \chi_{N_X}\}$ – множество научных журналов и конференций.

Образование малых команд

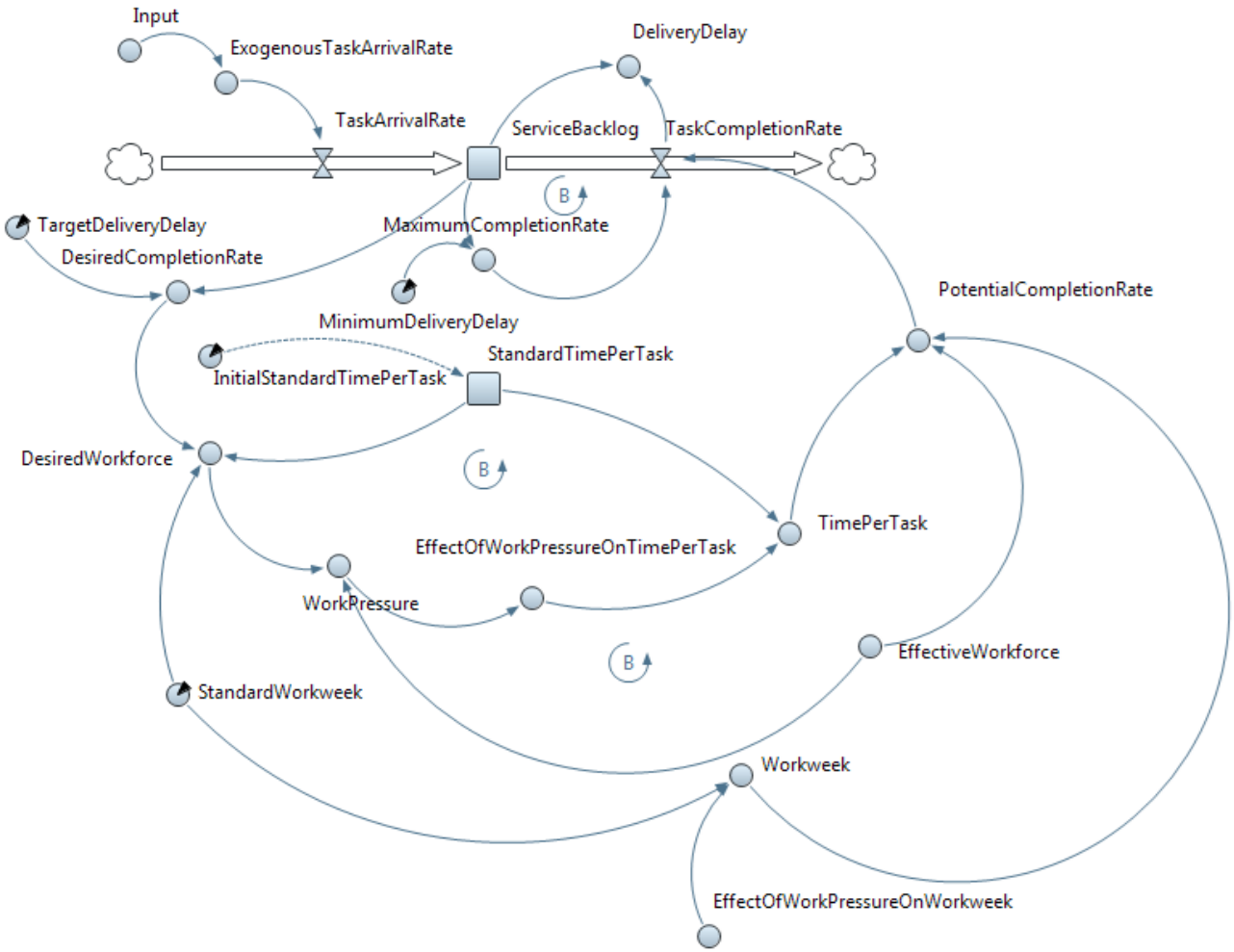




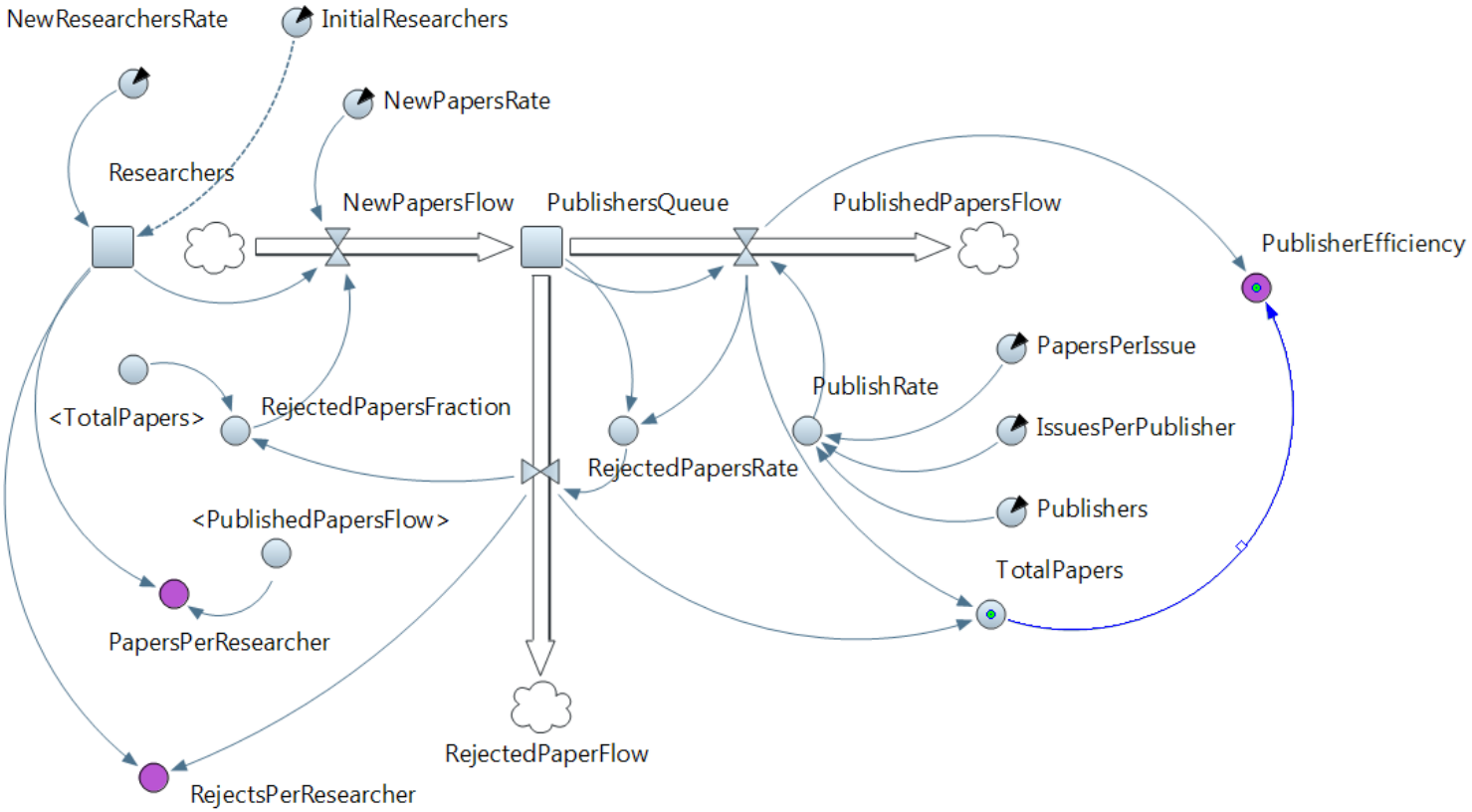
Модель персонала



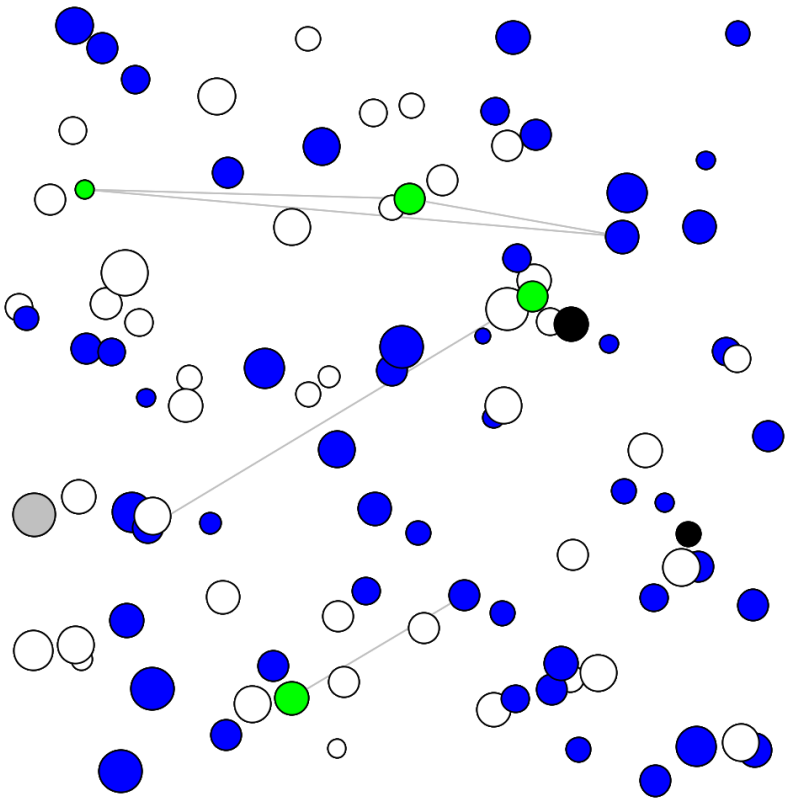
Модель выполнения заданий



Модель проведения исследований



Много агентное моделирование



Популяции

employees
Employee [100]

papers
Paper [266]

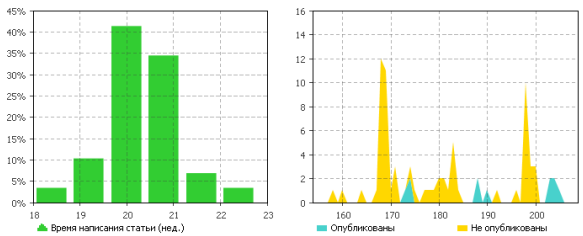
publishers
Publisher [11]

Тема конференции: ГИР; ГРР; Капитальное строительство; HSE;

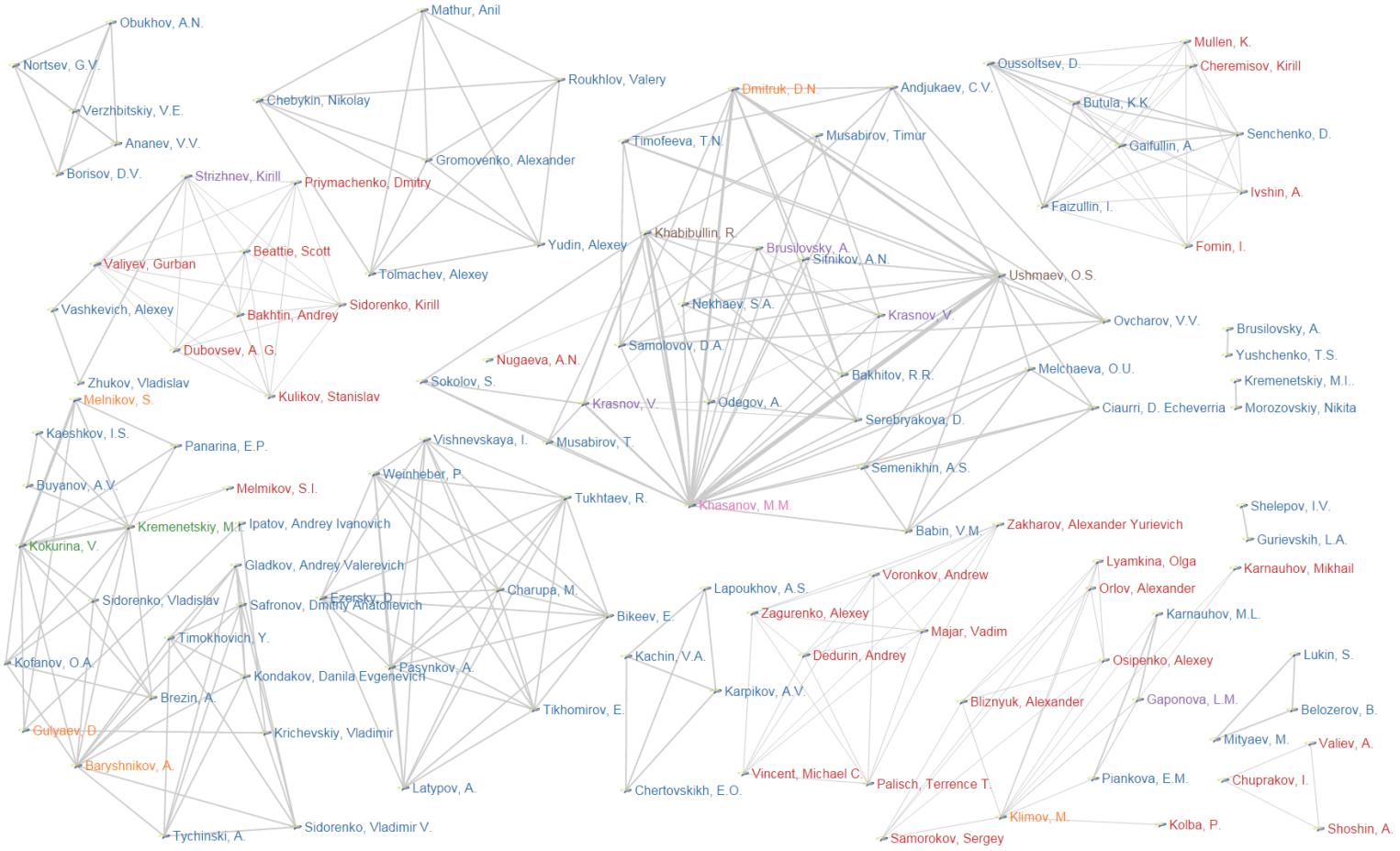
Модель процесса написания научных статей.
employees - сотрудники организации.
publishers - издатели, конференции.
papers - научные статьи.

Автор Соавтор Сотрудник "Не буду автором"

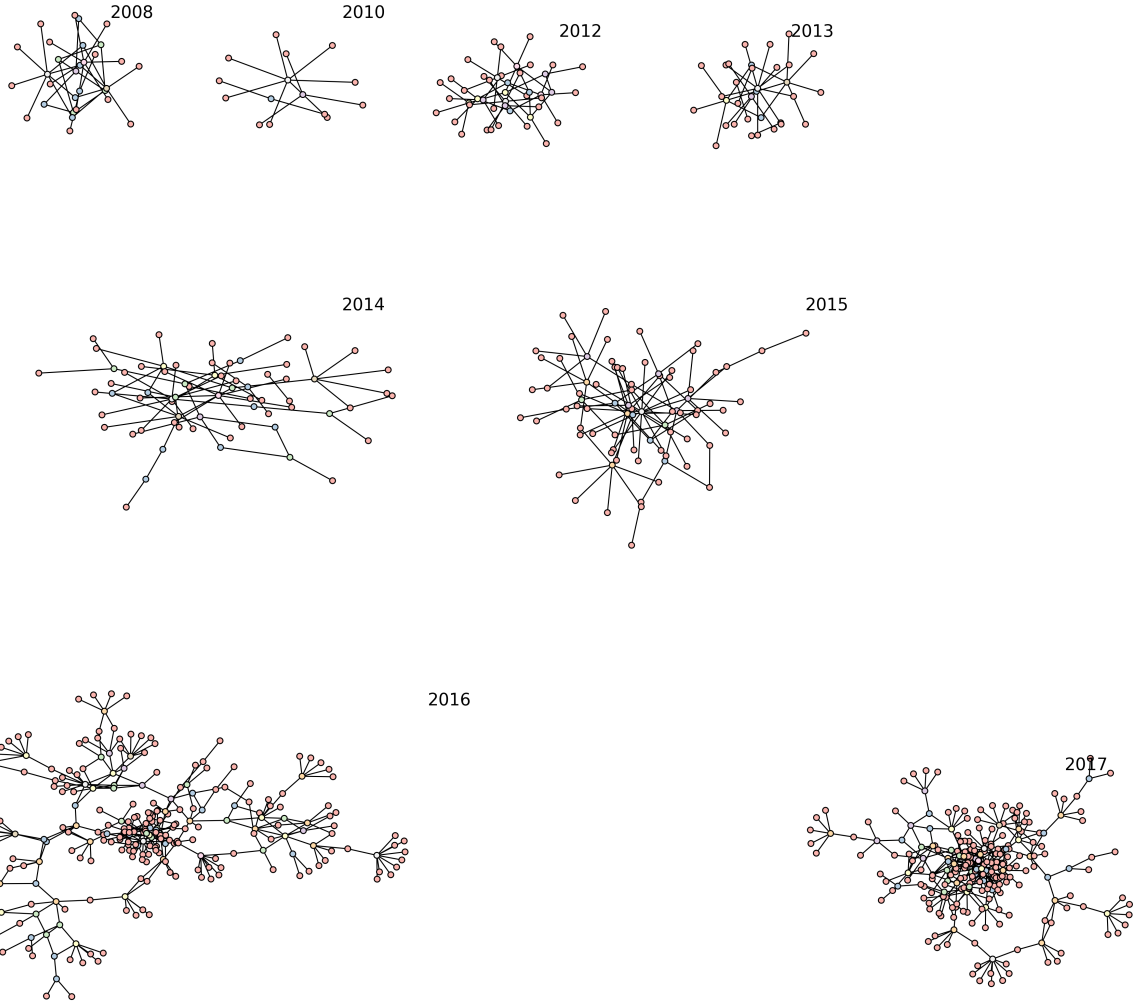
2018 Jan



Синтетический граф соавторства



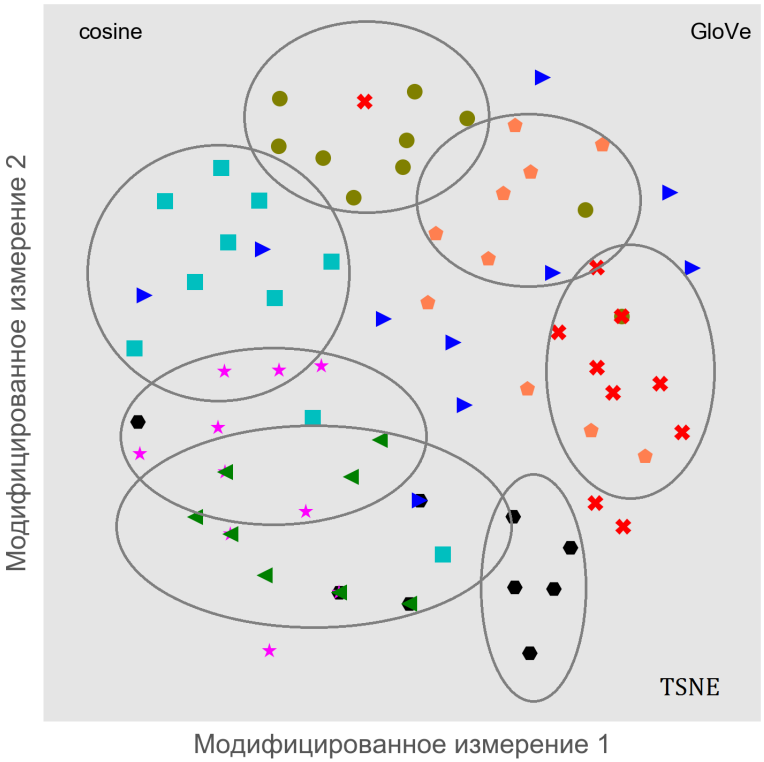
Кластеризация графа соавторства на основе алгоритма "обратного распада".

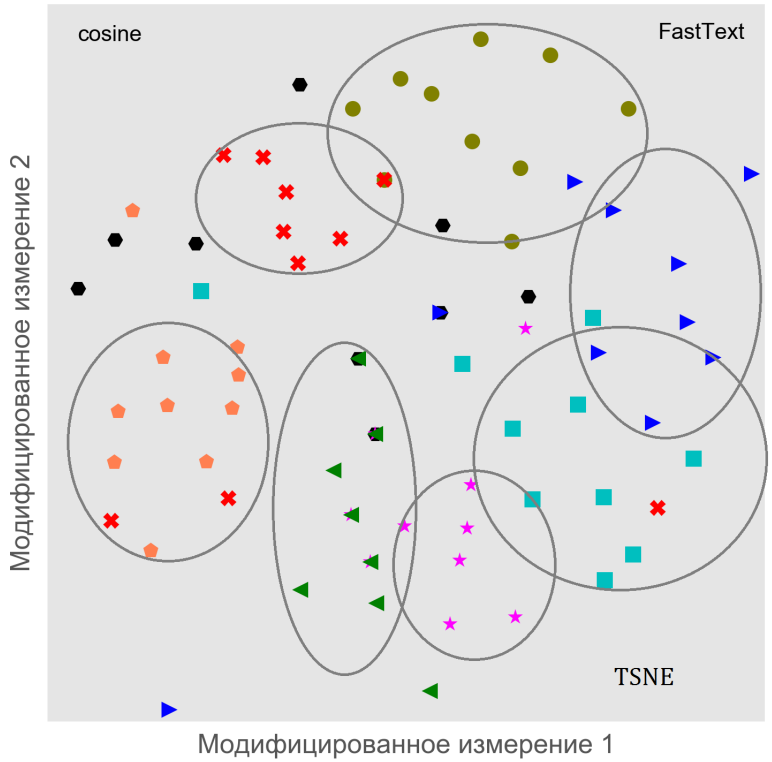


Тематическая модель текста (I)

1. Автор показал результативность подхода к улучшению интерпретируемости тематик на основе последовательной регуляризации.
2. Примененные методы управление отношением “плотность-разрежённость” открывают возможности настройки модели на предметную область текстов. Автор показал принципы создания и настройки модели тематик, которые позволяют вести интеллектуальный поиск (разведку) высоко сфокусированныхисточников знаний.
3. Кластеризация топиков была проверена с помощью двух методов для векторизации слов ($FastText$, $GloVe$) и двух методов для уменьшения размерности векторного пространства ($TSNE$, MDS). Результаты представлены в виде диаграмм и уверено показывают наличие кластеров.

Тематическая модель текста (II)





Тематики, выделенные с помощью PLSA с последовательной регуляризацией, образуют кластеры.

Практическая значимость результатов исследования

По теме исследования опубликовано более **40** работ; **15** из них опубликованы в рецензируемых научных изданиях, рекомендованных ВАК.
Доклады автора опубликованы в **6** сборниках из списка Web of Science.
Получены **3** свидетельства о регистрации программ для ЭВМ.

Результаты, полученные в ходе настоящего исследования – модели, методы, алгоритмы, комплексы программ были использованы для решения практических задач по оценке научной эффективности НТЦ в энергетической отрасли.