# ПОСТРОЕНИЕ И ИССЛЕДОВАНИЕ МАТЕМАТИЧЕСКОЙ МОДЕЛИ ВЕБ-ПРОСТРАНСТВА

Ю.И. Шокин*, А.Ю. Веснин**, А.А. Добрынин**, О.А. Клименко*,
Е.В. Рычкова*, М.Я. Филиппова***
*Институт вычислительных технологий СО РАН
**Институт математики им. С. Л. Соболева СО РАН
***Институт систем информатики им. А.П. Ершова СО РАН

**Аннотация:** В работе исследуется веб-пространство трех академических сообществ: Сибирского отделения РАН, Общества Фраунгофера в Германии и научных учреждений Республики Сербия. Объектами анализа являются сайты научных организаций и гиперссылки между ними, порождающие соответствующие веб-графы. Методами теории графов и вебометрики проводится сравнение численных и структурных характеристик веб-графов академических сообществ, строятся рейтинги сайтов.

**Ключевые слова:** веб-пространство, веб-граф, вебометрика, рейтинг сайтов

## Введение

В настоящее время задача изучения веб-пространства является актуальной в связи со стремительным развитием сети интернет и ростом объема представленных в ней ресурсов. Под веб-пространством понимаются сайты сети интернет, связанные друг с другом гиперссылками. Анализ свойств интернета как математического объекта впервые был начат в работах Р. Албертa и А.-Л. Барабаши [1]. Проблематика исследований включает поиск адекватных представлений интернета в виде сложной сетевой структуры и исследование ее свойств, нахождение математических параметров, характеризующих такую сеть, определение и предсказание изменений этих параметров при эволюции сети. Для изучения содержательных и логических связей между объектами интернета удобно использовать их представление в виде веб-графа. Как правило, при построении веб-графа в качестве вершин рассматриваются отдельные страницы сайтов или сайты, рассматриваемые как единое целое. В настоящей работе

под веб-графом понимается ориентированный граф, вершины которого соответствуют веб-сайтам, а отношение между сайтами определяется ссылками друг на друга.

Для изучения веб-пространства используются также методы вебометрики — современного раздела информатики, объектом изучения которого являются информационные ресурсы, структура и технологии интернета. Развитие этого направления началось в 1997 г. после работы Т. Алминда и П. Ингверсена [2]. Методы вебометрики позволяют оценивать эффективность и востребованность интернет-ресурсов с помощью доступных количественных показателей их деятельности. К таким показателям относятся, например, количество посещений, индексы цитируемости, степень связности определенных групп интернет-ресурсов, объемы доступных данных.

Одним из направлений исследований, представляющих интерес для организации и формирования научных интернет-сообществ, является анализ существующих веб-пространств, порождаемых сайтами университетов и академических научных организаций (см., например, [2–13]). В настоящей работе изучаются веб-пространства сайтов научных организаций Сибирского отделения Российской академии наук (СО РАН), научных организаций, объединенных в Общество Фраунгофера в Германии (ОФ), и научных учреждений, входящих в Сербскую академию наук и искусств и Zajednice instituta Srbije. Для указанных веб-пространств проведено сравнение их численных и структурных характеристик, для первых двух веб-пространств определены рейтинги входящих в них сайтов.

## 1. Анализ структуры веб-графов научных организаций методами теории графов

### 1.1. Веб-графы научных организаций

Одним из подходов к анализу структуры веб-пространств, порождаемых веб-сайтами и гиперсвязями между ними, является использование методов теории графов [14–16]. Для этого в качестве модели веб-пространства используется веб-граф, в котором вершины соответствуют сайтам, а оотношение между сайтами задается наличием ссылок между ними. Дуга графа выходит из вершины $v$ и входит в вершину $u$, если сайт, соответствующий вершине $v$, содержит хотя бы одну ссылку на страницы сайта, соответствующего вершине $u$. Количество ссылок с одного сайта на другой

задает вес соответствующей дуги (сссылки сайта на себя не учитываются). Таким образом, веб-граф является ориентированным графом, в котором любая пара вершин может быть соединена одной дугой или двумя противоположно направленными дугами. Число дуг в кратчайшем ориентированном пути между парой вершин равно наименьшему числу шагов при переходе по ссылкам с одного сайта на другой. При изображении графов и их фрагментов пара противоположно направленных дуг для наглядности будет заменяться одной дугой с двумя стрелками на концах.

Веб-графы научных организаций Общества Фраунгофера, СО РАН и научных организаций, входящих в Сербскую академию наук и искусств и Zajednice instituta Srbije, обозначим через *G*, *R* и *S* соответственно.

Веб-граф *G*, изображенный на рис. 1, содержит 72 вершины и 321 дугу и отражает связи научных организаций Общества Фраунгофера по состоянию на 8 апреля 2013 г. Список организаций ОФ представлены в [17].

Веб-граф *R*, содержащий 95 вершин и 949 дуг, соответствует состоянию веб-пространства научных организаций СО РАН на 29 октября 2013 г. В этот граф включены все научные организации из информационной системы «Организации и сотрудники СО РАН» [18]. Структура графа *R* показана на рис. 2 [19].

Веб-граф *S* на рис. 3 имеет 59 вершин и 106 дуг и отражает состояние веб-пространства научных организаций Республики Сербия и Zajednice instituta Srbije на 7 апреля 2013 г. Список организаций, соответствующих вершинам графа *S*, дан в [20].

На рис. 4 приведен пример визуализации рассматриваемых графов при радиально-осевой укладке, наглядно выявляющий различия в их структуре.

### 1.2. Численные характеристики структуры графов

Численная характеризация структуры графа, проведенная на основе анализа ее локальных фрагментов, полезна при изучении графов большого размера, так как часто не требует трудоемких расчетов. Для описания тех или иных структурных особенностей графов используют инварианты графов, которые, как правило, являются функциями, ставящими в соответствие графу некоторое число. Значение инвариантов зависит только от структуры графа, т.е. на изоморфных графах инвариант всегда

принимает одинаковые значения. Далее рассмотрим инварианты, отражающие вклад вершин, дуг и окрестностей вершин графа в формирование его структуры.

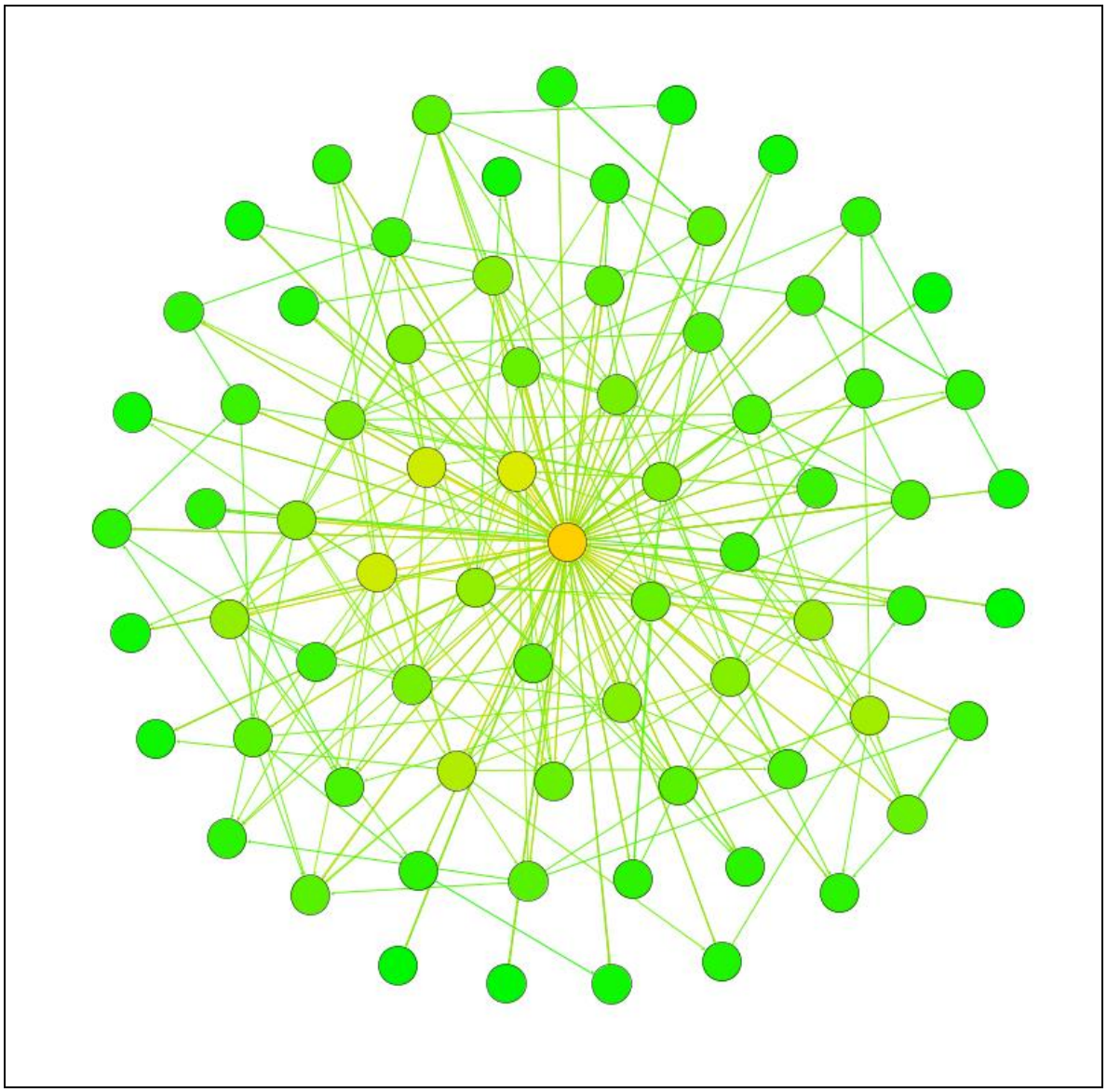


**Рис. 1. Веб-граф *G* научных организаций Общества Фраунгофера.**

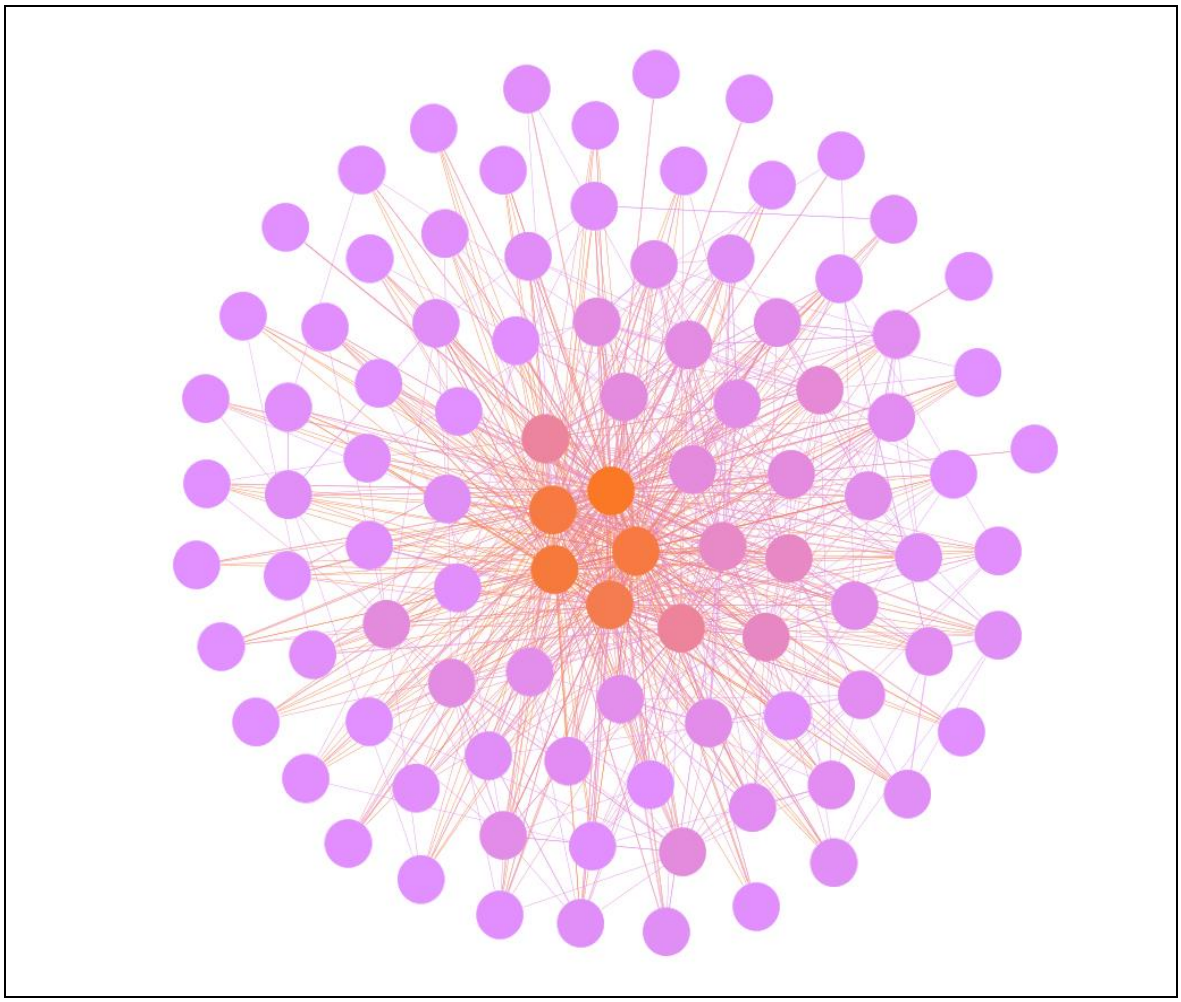


**Рис. 2. Веб-граф *R* научных организаций СО РАН.**

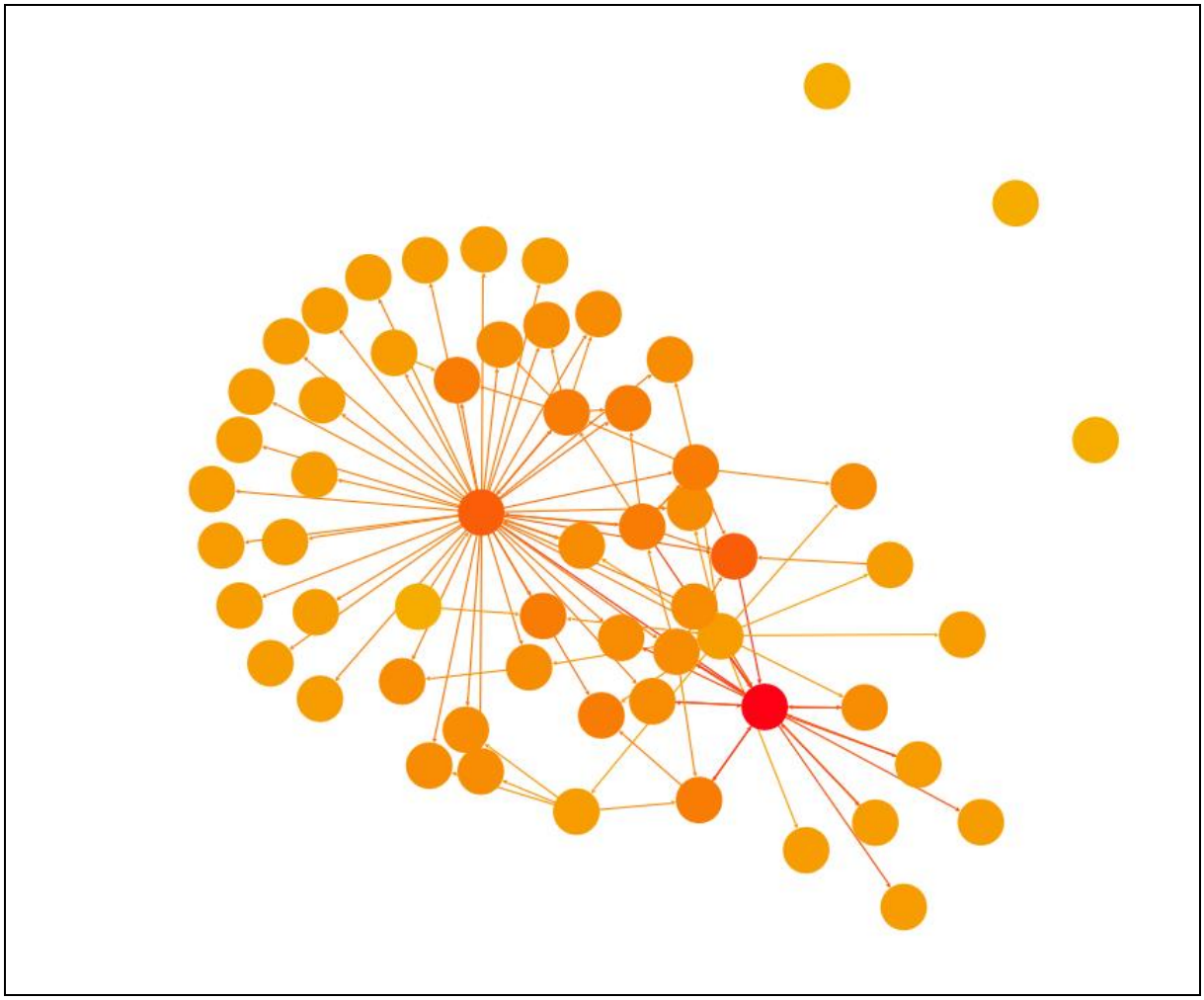


**Рис. 3. Веб-граф *S* сербских научных организаций.**

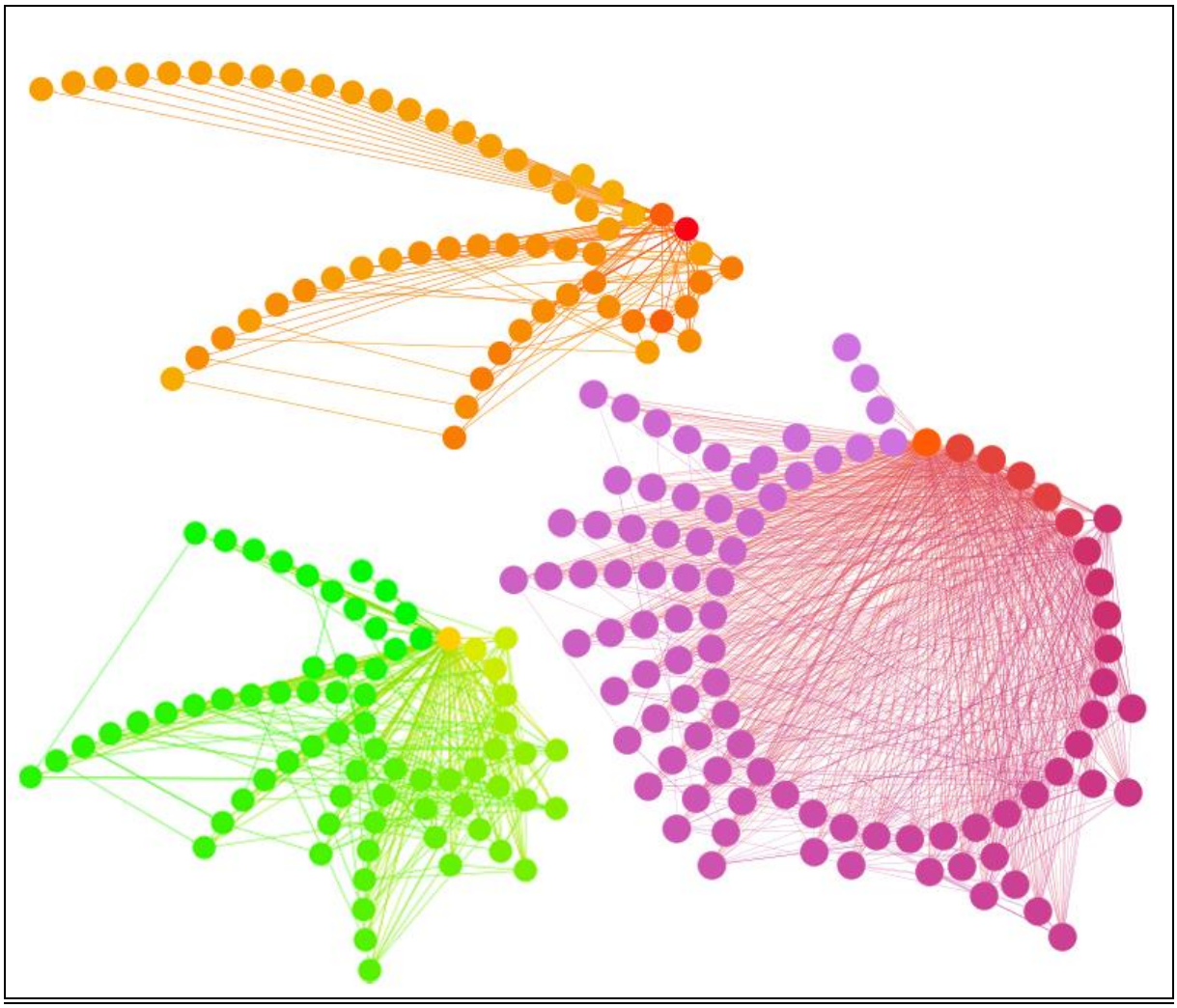


**Рис. 4. Сравнительные изображения графов *S* (оранжевый цвет), *G* (зеленый цвет) и *R* (фиолетовый цвет).**

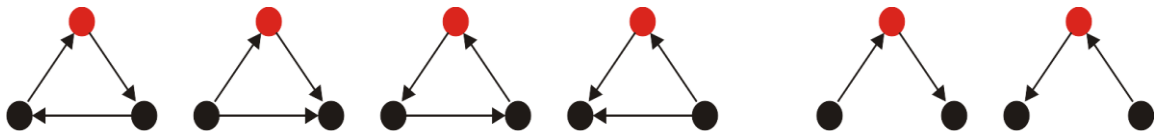
Индекс вершин графа $c_v(H)$. Этот параметр показывает какая часть сайтов веб-графа включена в информационное взаимодействие с другими сайтами, хотя бы попарное. Пусть ориентированный граф $H$ имеет $n$ вершин, и $k$ из них имеют хотя бы одну исходящую или входящую дугу. *Индексом вершин* в графе $H$ называется величина $c_v(H) = k/n$. Близость $c_v(H)$ к нулю говорит о том, что имеется значительное количество изолированных сайтов, то есть таких, которые не связаны с другими сайтами. Это может иметь место, например, в начальной стадии формирования веб-пространства. Равенство $c_v(H) = 1$ означает, что каждая организация вовлечена в информационное взаимодействие с другими на уровне веб-сайтов.

Индекс дуг графа $c_a(H)$. Максимальное число дуг в ориентированном графе $H$ с $n \geq 2$ вершинами равно $n(n-1)$. Пусть число дуг в графе $H$ равно $t$. *Индексом дуг* в графе $H$ называется величина $c_a(H) = t / n(n-1)$. В [21] эта величина называется плотностью сети. Индекс дуг показывает какая часть дуг графа участвует в установлении информационного взаимодействия между сайтами. Максимальное значение, $c_a(H) = 1$, достигается когда любые две вершины графа $H$ соединены парой противоположно направленных дуг. В этом случае все сайты ссылаются друг на друга, обеспечивая кратчайший переход с одного сайта на другой.

Коэффициент кластеризации графа $cc(H)$. Под *окрестностью* вершины $v$ будем понимать множество вершин графа, соединенных с $v$ дугами без учета их ориентации. Пусть $V_2$ есть множество вершин ориентированного графа $H$, окрестность которых содержит не менее чем две вершины. Для вершины $v$ графа $H$ обозначим через $H_v$ ориентированный подграф, порожденный окрестностью вершины $v$. *Коэффициентом кластеризации вершины* $v$ называется величина $c_a(H_v)$, т.е. индекс дуг подграфа, порожденного окрестностью вершины $v$ [22]. *Коэффициент кластеризации графа* $H$ определим как $cc(H) = \frac{1}{|V_2|} \sum_{v \in V_2} c_a(H_v)$. Таким образом, $cc(H)$ показывает как в среднем заполнены дугами окрестности вершин графа.

Коэффициент транзитивности графа $\tau(H)$. Рассмотрим в графе ориентированные пути длины 2, центральная вершина которых изображена на рис. 5 красным цветом. Обозначим через $N_\Delta$ число всех ориентированных путей длины 2 в графе $H$ таких, что концевые вершины $u$ и $v$ этих путей соединены дугой без учета ориентации (все

возможные конфигурации показаны на рис. 5а). Число всех ориентированных путей длины 2, между концевыми вершинами которых нет дуг, обозначим через $N_\Lambda$ (все конфигурации приведены на рис. 5б). *Коэффициент транзитивности* $\tau(H)$ ориентированного графа $H$ определим как $\tau(H) = N_\Delta / N_\Lambda$ [23].



*а*) *б*)

**Рис. 5. Конфигурации для вычисления коэффициента транзитивности.**

Диаметр графа *diam*(*H*). Расстояние между парой вершин в ориентированном графе *H* определяется как наименьший по числу дуг путь, соединяющий эти вершины, причем все вершины этого пути разные. Диаметр графа *H* определяется как наибольшее расстояние между парами вершин в графе. Таким образом, диаметр веб-графа показывает, какое наибольшее число шагов можно сделать по уникальным сайтам, переходя по ссылкам с сайта на сайт.

Значения введенных выше численных параметров для веб-графов научных организаций *G*, *R* и *S* представлены в таблице 1.

Таблица 1. Характеристики веб-графов научных организаций.

| **Граф** | $c_v$ | $c_a$ | ***cc*** | **τ** | ***diam*** |
|---|---|---|---|---|---|
| *G* | 1,00 | 0,06 | 0,09 | 0,10 | 2 |
| *R* | 0,98 | 0,11 | 0,07 | 0,24 | 3 |
| *S* | 0,95 | 0,03 | 0,03 | 0,07 | 6 |

Значения индекса вершин $c_v$ показывают, что в веб-пространстве научных организаций ОФ все сайты включены в информационное взаимодействие, в то время как некоторые сайты СО РАН и сербских научных организаций не связаны с другими. Значения индекса дуг $c_a$ для графа *R* почти в два-три раза превышает значения для

графов *G* и *S*. Значения коэффициента кластеризации *cc* показывают, что в среднем заполнение дугами окрестности вершин во всех графах мало. Величины коэффициента транзитивности для графа *R* заметно выше, чем для других графов. Малый диаметр графа ОФ объясняется наличием вершины с почти максимально возможными полустепенями. Из этой вершины выходят дуги до всех других вершин, и в нее заходят дуги также из всех вершин, кроме одной.

### 1.3. Входящие и исходящие связи вершин графов

Естественными характеристиками вершины *v* ориентированного графа являются число исходящих из нее дуг $deg_+(v)$ (полустепень исхода) и число входящих в нее дуг $deg_-(v)$ (полустепень захода). Вершина *v*, для которой $deg_+(v) = deg_-(v) = 0$, называется изолированной. На рис. 6 и 7 приводятся графики функции распределения числа вершин веб-графов *G*, *R* и *S* по их полустепеням исхода и захода. Функция распределения в точке *k* горизонтальной оси равна числу вершин с $deg_\pm(v) \le k$.

Средние полустепени исхода/захода вершин в рассматриваемых графах равны $avr(G) = 4{,}46$, $avr(R) = 9{,}99$ и $avr(S) = 1{,}8$ (суммы полустепеней исхода и захода всех вершин графа всегда равны).

У трех графов наблюдается сильное различие в числе вершин, из которых не выходит ни одной дуги (1, 21 и 38 вершин). В графе *G* полустепень исхода у почти всех вершин ограничена значением 10, а в графе *S* – значением 6. В графах *G* и *R* из вершин с максимальной полустепенью исхода дуги идут почти ко всем вершинам графов, в то время как в графе *S* из подобной вершины можно перейти только в 73% вершин.

Число вершин, в которые не входит ни одна дуга, во всех графах невелико (0, 2 и 3 изолированные вершины). Полустепени захода вершин в графе *S* ограничены значением 9, а подавляющее число вершин имеет полустепень захода не более 3. В графе *G* полустепени захода почти всех вершин тоже ограничены значением 9. В графах *G* и *R* существуют вершины, в которые заходят дуги из почти всех других вершин.

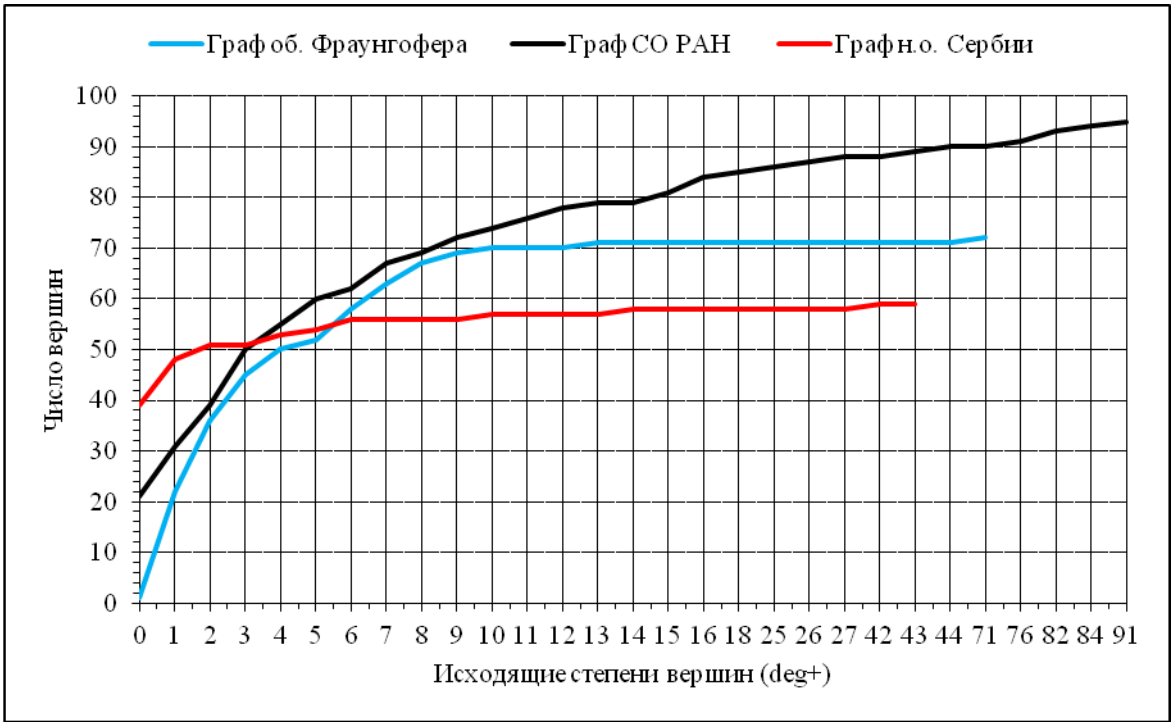



**Рис. 6. Распределение вершин в графах *G*, *R* и *S* по полустепеням исхода $deg_+$.**

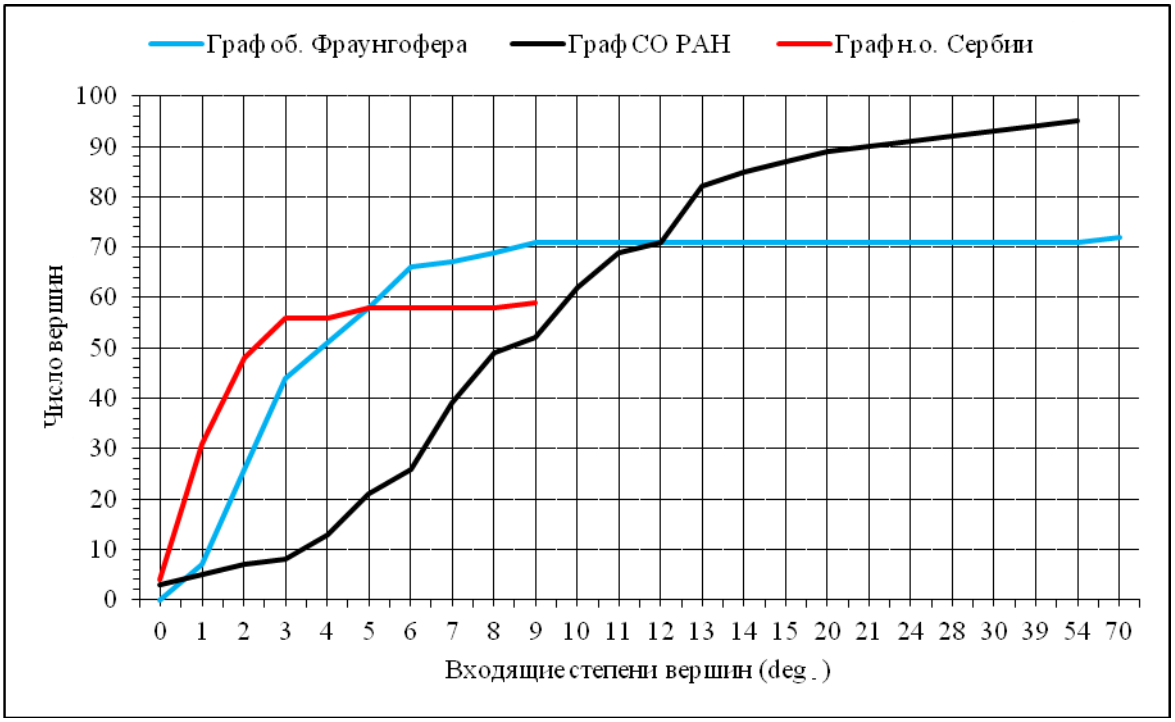



**Рис. 7. Распределение вершин в графах *G*, *R* и *S* по полустепеням захода $deg_-$.**

### 1.4. Веб-коммуникаторы в графах

Доступность и использование информационных ресурсов веб-пространства можно характеризовать соотношением между полустепенями исхода и захода вершин веб-графа. Большие полустепени вершины обеспечивают тесные связи соответствующего сайта с остальным веб-пространством. Выделим три типа возможного соотношения числа входящих и исходящих дуг (рис. 8). Вершины первого типа называют *индукторами* (мало входящих дуг, много исходящих), второго типа — *коллекторами* (много входящих дуг, мало исходящих), а третьего типа — *посредниками* (много как входящих, так и исходящих дуг). Такие типы вершин образуют множество *веб-коммуникаторов* графа.

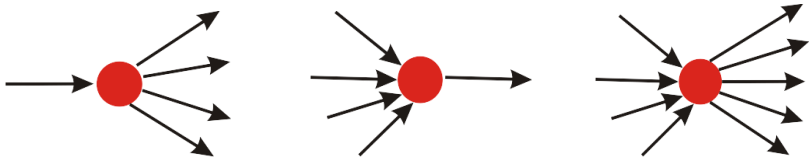
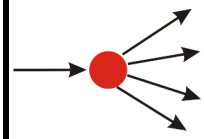
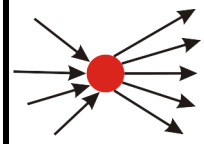
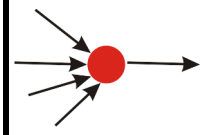


**Рис. 8. Веб-коммуникаторы: индуктор, коллектор и посредник.**

Коллекторы могут соответствовать сайтам организаций, в которых происходит накопление, хранение и обработка данных. Это могут быть библиотеки, хранилища данных, центры коллективного пользования и обработки данных, справочные ресурсы, журналы. Посредниками могут быть вершины, соответствующие головным сайтам в какой-то области науки, порталам научных центров, сайтам институтов с высокой степенью научной кооперации, официальным сайтам. Индукторами могут являться сайты недавно созданных организаций.

Для отнесения вершины графа к веб-коммуникаторам того или иного типа используем численные параметры, характеризующие соотношение между ее полустепенями. Пороговые значения параметров можно задавать в зависимости от распределения полустепеней вершин в графе. Будем считать, что малые полустепени в веб-коммуникаторах должны быть меньше средних полустепеней, в большие полустепени должны превышать средние полустепени.

Таблица 2. Веб-коммуникаторы в графах $G$, $R$ и $S$ при $\Delta_i = \Delta_c = \Delta_t = 2$.

| | **Общество Фраунгофера** $avr(G) = 4.5$ | | **Сибирское отделение** $avr(G) = 9.9$ | | **Научные орг. Сербии** $avr(G) = 1.8$ | |
|---|---|---|---|---|---|---|
| | ***Количество*** | **($deg_-$,$deg_+$)** | ***количество*** | **($deg_-$,$deg_+$)** | ***количество*** | **($deg_-$,$deg_+$)** |
| Индуктор  | 7% (5) | (1,6) | 1% (1) | (2,43) | 3% (2) | (4,1) |
| | | (1,7) | | - | | (14,1) |
| | | (2,8) | | - | | - |
| | | (2,10) | | - | | - |
| | | (3,13) | | - | | - |
| Посредник  | 7% (5) | (6,6) | 7% (7) | (10,10) | 8% (5) | (2,2) |
| | | (6,7) | | (11,11) | | (2,3) |
| | | (7,9) | | (13,12) | | (4,2) |
| | | (9,8) | | (15,13) | | (5,3) |
| | | (70,71) | | (14,15) | | (10,11) |
| Коллектор  | 10% (7) | (5,1) | 6% (6) | (11,1) | 2% (1) | (1,5) |
| | | (6,1) | | (11,2) | | - |
| | | (8,1) | | (10,2) | | - |
| | | (9,2) | | (13,1) | | - |
| | | | | (13,2) | | - |

Более точно, используем следующие правила для определения веб-коммуникаторов: вершина $v$ в графе $H$ является

– индуктором, если $deg_-(v) < avr(H)$, $deg_+(v) > avr(H)$ и $deg_+(v) / deg_-(v) > \Delta_i$;

– коллектором, если $deg_-(v) > avr(H)$, $deg_+(v) < avr(H)$ и $deg_-(v) / deg_+(v) > \Delta_c$;

– посредником, если $deg_-(v) > avr(H)$, $deg_+(v) > avr(H)$ и $|deg_+(v) - deg_-(v)| \leq \Delta_t$,

где $avr(H)$ — средняя степень вершин в графе $H$, а $\Delta_i$, $\Delta_c$ и $\Delta_t$ — заданные границы.

В таблице 2 приводятся веб-коммуникаторы в веб-графах научных организаций ОФ, СО РАН и Сербии при $\Delta_i = \Delta_c = \Delta_t = 2$. Средние степени вершин указаны в верхней части таблицы. Количество коммуникаторов дано в процентах от числа вершин, рядом в скобках указаны абсолютные значения. В столбцах также приводятся значения полустепеней веб-коммуникаторов (различающиеся пары).

В множество веб-коммуникаторов попадают 25% вершин в графе $G$, в то время как в графах $R$ и $S$ число таких вершин составляет 14% и 13% соответственно. При $\Delta_t = 2$ вершины с большими полустепенями не попадают в множество посредников.

Например, вершина с полустепенями (91,54), соответствующая сайту ПОРТАЛ СО РАН, не будет посредником, хотя по своему положению в веб-пространстве СО РАН является им. Если при нахождении веб-коммуникаторов не налагать ограничений на разницу между полустепенями вершин (положить $\Delta_t = 100$), то графы *G, R* и *S* будут иметь 7% (5 вершин), 21% (20) и 14% (8) посредников.

### 1.5. Агрегированное представление структуры графов

Для описания структуры больших веб-пространств часто используется их представление в виде модели «галстук-бабочка» [24]. В веб-графе выделяется максимальная сильно связная компонента, по отношению к которой классифицируются остальные вершины графа. Подграф называется сильно связной компонентой графа, если между любой парой его вершин существует ориентированный путь. Проходя по ссылкам сайтов, вершины которых попали в такую компоненту, можно всегда посетить любую вершину компоненты. На рис. 9 показано как выглядит веб-граф в таком представлении. Его центральную часть образует максимальная сильно связная компонента (SCC). Левая часть (IN) состоит из вершин, пути из которых ведут в эту компоненту. Правую часть (OUT) образуют вершины, в которые ведут пути из компоненты SCC. Вершины подграфов, называемых отростками (tendrils), связаны путями только с множествами вершин IN и OUT. На рис. 9 эти пути изображены в виде толстых стрелок, выходящих из IN или входящих в OUT. Некоторые вершины из множества IN могут быть связаны с вершинами из OUT путями, избегающими сильно связную компоненту SCC. На рис. 9 такие подграфы, называемые туннелями (tubes), изображены в виде толстой дуги из IN в OUT. Другие подграфы, вершины которых не связанны путями с описанными выше частями графа, образуют отдельные компоненты (показаны внизу на рис. 9).

На рис. 10 показана структура веб-графов научных организаций Сербии, СО РАН и общества Фраунгофера в виде модели «галстук-бабочка». В процентах указано количество вершин в частях графа. В графах организаций Сербии и СО РАН несвязные компоненты образованы изолированными вершинами. В графе Сербии есть один туннель. У общества Фраунгофера левая часть IN не содержит вершин.

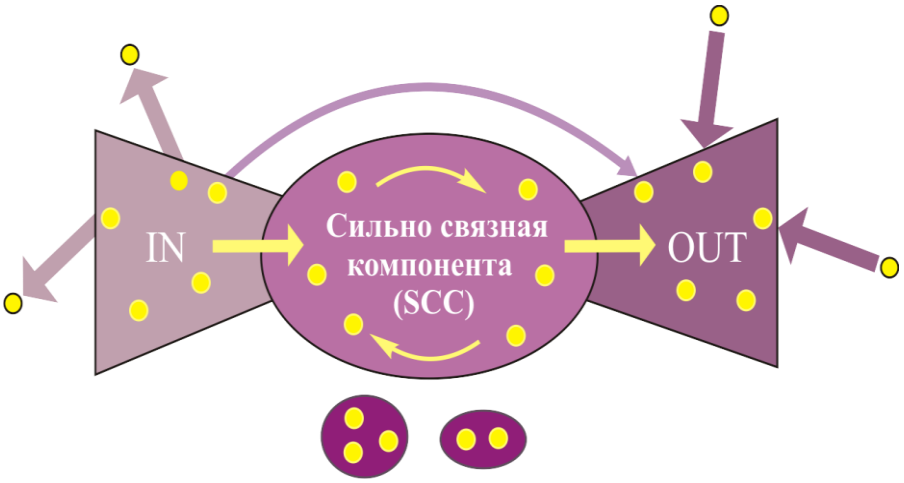



**Рис. 9. Представление веб-графа в модели «галстук-бабочка».**

Напомним, что центральная часть бабочки имеет полезное свойство – из любого сайта в SCC всегда можно достичь информационных ресурсов другого сайта этой части. Размер множества SCC в веб-графах рассматриваемых академических сообществ увеличивается с 29% до 99% всех вершин в графах *S* и *G*. Большой размер множества SCC в графе общества Фраунгофера обеспечивается существованием портала, из которого выходят дуги во все другие сайты, и в него также входят дуги из почти всех вершин (кроме одной).

Небольшой размер веб-графа научных организаций Сербии позволяет наглядно показать расположение его частей в модели «галстук-бабочка» (рис. 11). Вершины сильно связной компоненты, SCC, изображены красным цветом (17 вершин), вершины множества OUT выделены голубым цветом (38 вершин), а единственная вершина части IN имеет черный цвет. Из нее выходят две черные дуги – одна в красную вершину из SCC, а другая – в голубую вершину из OUT (туннель). Три изолированные вершины образуют несвязанные компоненты (выделены фиолетовым цветом).


граф *S* организаций республики Сербия


граф *R* Сибирского отделения РАН


граф *G* Общества Фраунгофера

**Рис. 10. Представление структуры веб-графов *S*, *R* и *G*.**

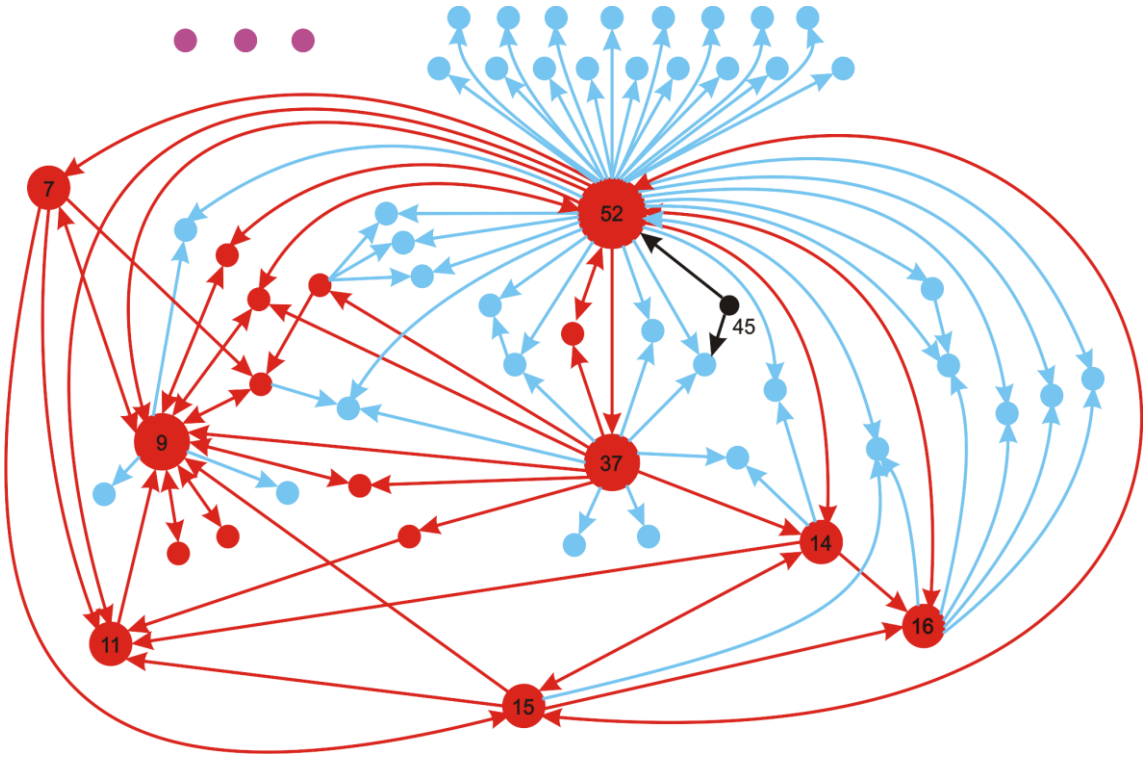



**Рис. 11. Части веб-графа *S* научных организаций Республики Сербия. 7 — Institute of Technical Sciences, 9 — Serbian Academy of Sciences and Arts, 11 — Institute for Biological Research “Siniša Stanković”, 15 — Vinča Institute of Nuclear Sciences, 16 — Institute of Chemistry, Technology and Metallurgy, 37 — Institute of Economic Sciences, 45 — Institute of Agricultural Economics, 52 — Zajednice instituta Srbije.**

### 1.6. Группы наиболее тесно связанных организаций

Представляет интерес выявление в веб-графах подмножеств вершин, которые могут быть достижимы друг из друга за небольшое число шагов.

Основанием ориентированного графа будем называть неориентированный граф с тем же множеством вершин, в котором пара вершин соединена неориентированным ребром, если эти вершины были соединены в исходном графе одной дугой или двумя противоположно направленными дугами. Таким образом, ребро основания ориентированного графа отражает факт наличия ссылок между парой соответствующих сайтов без указания направления ссылок.

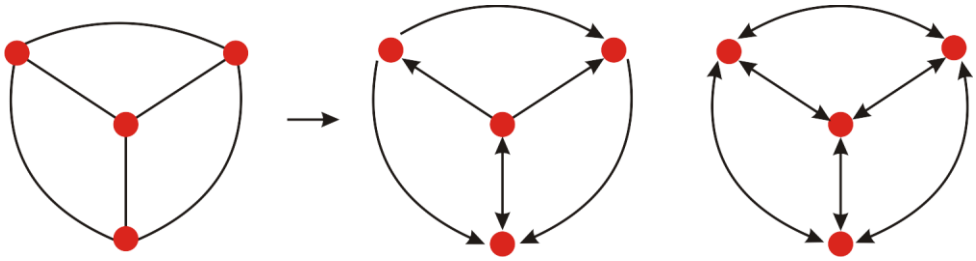
Рассмотрим типы подграфов, которые могут образовывать группы тесно связанных друг с другом организаций. Любая пара вершин в таких подграфах соединена хотя бы одной дугой, а типы подграфов определяются ориентацией дуг.

Неориентированный подграф называется *полным*, если любая пара его вершин соединена ребром. Полный подграф называется *кликой* графа, если он не содержится ни в каком другом полном подграфе, т.е. является максимальным по включению (такие подграфы называют еще максимальными кликами). Клику с числом вершин $k$ будем называть *k-кликой*. Клики основания графа $H$ порождают в $H$ ориентированные подграфы, которые будем называть ориентированными кликами. В зависимости от ориентации дуг выделим два типа ориентированных клик.

1). Ориентированная клика называется *компактной*, если каждая пара ее вершин соединена двумя противоположно направленными дугами. Компактная клика представляет собой оптимальный фрагмент с точки зрения быстроты доступа к информационным ресурсам ее вершин. Эволюция веб-графов организаций в какой-либо узкой области науки может приводить к компактной клике.

2). Ориентированная клика называется *сильной,* если она является сильно связной компонентой графа, но не является компактной кликой. Таким образом, в сильной клике все вершины достижимы друг из друга, но между некоторыми вершинами расстояние больше, чем в случае компактной клики.

Примеры клик указанных типов приведены на рис. 12. Ясно, что добавляя новые дуги, всегда можно перейти от сильной клики к компактной.



*а*) *б*) *в*)

**Рис. 12. Полный подграф (*а*), сильная (*б*) и компактная (*в*) клики.**

Клики веб-графа образуют близкие группы сайтов. Если в клику добавляются новые дуги, то сильная клика может стать компактной. Это может быть использовано для ускорения доступа к информационным ресурсам внутри группы. В таблице 3 приводятся распределения числа подграфов разных типов в графах *R, G* и *S*.

Таблица 3. Распределение клик графов по числу вершин.

| **граф** | **Число вершин →** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *R* | полные подграфы | 2 | 3 | 4 | 12 | 21 | 21 | 55 | 48 | 30 | 7 |
| | сильные клики | - | 2 | - | 6 | 13 | 16 | 43 | 43 | 30 | 7 |
| | компактные клики | - | - | - | - | - | - | - | - | - | - |
| | компактные подграфы | 19 | 24 | 11 | 13 | 10 | 4 | | | | |
| *G* | полные подграфы | 4 | 83 | 31 | 2 | 1 | - | - | - | - | - |
| | сильные клики | - | 76 | 30 | 2 | 1 | - | - | - | - | - |
| | компактные клики | 4 | 5 | - | - | - | - | - | - | - | - |
| | компактные подграфы | 48 | 14 | - | - | - | - | - | - | - | - |
| *S* | полные подграфы | 34 | 19 | 4 | 1 | - | - | - | - | - | - |
| | сильные клики | - | 2 | 2 | 1 | - | - | - | - | - | - |
| | компактные клики | - | 2 | - | - | - | - | - | - | - | - |
| | компактные подграфы | - | 11 | - | - | - | - | - | - | - | - |

В графе СО РАН группы близких вершин имеют максимальную мощность 11, в то время как в двух других графах размер таких групп не превышает 6. Разницу между числом полных подграфов и сильных и компактных клик образуют подграфы, в которых есть вершины, не достижимые друг из друга. Компактные клики в графе *R* отсутствуют, а в графах *G* и *S* такие подграфы порождаются некоторыми парами и тройками вершин.

Кроме указанного подхода представляет интерес нахождение максимальных компактных подграфов без использования основания графа. Если в исходном графе удалить все дуги, которые не входят в контуры длины 2, то все клики в полученном графе будут компактными. В таблице 3 приводится количество максимальных

компактных подграфов в графах *R, G* и *S*. Такие фрагменты графов *G* и *S* имеют не более трех вершин, в то время как размер максимальных компактных подграфов в *R* достигает 7 вершин.

## 2. Применение методов вебометрики для определения рейтинга научной организации

Начиная с 2008 г., в Институте вычислительных технологий СО РАН строятся рейтинги сайтов научных организаций Сибирского отделения РАН [5,7,12]. Для нахождения рейтингов используется методика из [25]. По этой методике для оценки сайтов используются следующие параметры.

**V** — видимость сайта. Его значение равно количеству внешних ссылок с других сайтов на данный ресурс. Этот параметр вычисляется посредством усреднения количества внешних ссылок, найденных с помощью поисковых систем Яндекс [26], Google [27] и Bing [28]:

$$\mathbf{V} = [\mathbf{V}_{\text{Яндекс}} + \mathbf{V}_{\text{Google}} + \mathbf{V}_{\text{Bing}}]/3.$$

**S** — размер сайта. Значение **S** равно количеству веб-страниц сайта, определяемому поисковыми системами. Так как поисковые системы не всегда корректно определяют количество веб-страниц, то значение данного параметра может отличаться от реального размера сайта. Параметр **S** вычисляется усреднением значений размера сайта, получаемых с помощью указанных выше поисковых систем:

$$\mathbf{S} = [\mathbf{S}_{\text{Яндекс}} + \mathbf{S}_{\text{Google}} + \mathbf{S}_{\text{Bing}}]/3.$$

**R** — насыщенность сайта. Этот параметр определяется как суммарное количество файлов форматов Adobe Acrobat (.pdf), Microsoft Word (.doc) и Microsoft Powerpoint (.ppt), обнаруженных на сайте поисковыми системами:

$$\mathbf{R} = [\mathbf{R}_{\text{Яндекс}} + \mathbf{R}_{\text{Google}}]/2.$$

**Sc —** индекс цитирования, полученный из систем Индекс цитирования Яндекса [29] и Google Scholar [30]. Этот параметр является мерой значимости сайта.

Определение рейтинга сайтов научных организаций включает следующие этапы.

1. Вычисление значений параметров видимости **V**, размера **S** и насыщенности **R** для каждого исследуемого сайта.

2. Ранжирование значений параметров **V**, **S**, **R**. Массив значений параметра, например **V**, для всех сайтов упорядочивается по убыванию. Сайту, имеющему максимальное значение **V**, присваивается ранг $\mathbf{V}_r = 1$. Сайтам с одинаковыми значениями **V** присваиваются одинаковые ранги. Аналогичным образом вычислялись ранги $\mathbf{S}_r$ и $\mathbf{R}_r$ параметров **S** и **R**.

3. Вычисление ранга $\mathbf{Sc}_r$ индекса цитирования **Sc**. Вначале независимо вычисляются ранги для $\mathbf{Sc}_{\text{Яндекс}}$ и $\mathbf{Sc}_{\text{Google}}$. Затем для каждого сайта полученные ранги суммируются, и величина $\mathbf{Sc}_r$ определяется ранжированием этих сумм. Сайт с наименьшей суммой получает ранг $\mathbf{Sc}_r = 1$.

4. Суммирование определённых выше рангов для каждого исследуемого сайта:

$$\mathbf{W} = \mathbf{V}_r + \mathbf{S}_r + \mathbf{R}_r + \mathbf{Sc}_r.$$

5. Рейтинг сайтов формируется упорядочением значений **W** по возрастанию. Таким образом, итоговый ранг (позиция в текущем рейтинге) будет тем выше, чем меньше значение **W**. Сайтам с одинаковыми значениями **W** присваиваются одинаковые рейтинги.

В таблице 4 приводятся сайты научных организаций СО РАН, находящиеся на первых 15 местах в рейтинге (в столбце **Sc** приведено значение только из Google Scholar). Рейтинг всех 95 сайтов веб-пространства СО РАН представлен в [31]. Аналогично, в таблице 5 указаны первые 15 сайтов из рейтинга сайтов научных организаций Общества Фраунгофера. Рейтинг всех 72 сайтов веб-пространства ОФ приводится в [32].

Таблица 4. Рейтинг сайтов организаций СО РАН

| Рейтинг | Организация | Адрес сайта | V | S | R | Sc |
|---|---|---|---|---|---|---|
| 1 | Портал СО РАН | www.sbras.ru | 54863 | 73363 | 10438 | 620 |
| 2 | Институт вычислительных технологий СО РАН | www.ict.nsc.ru | 68067 | 107935 | 795 | 154 |
| 2 | Институт цитологии и генетики СО РАН | www.bionet.nsc.ru | 6046 | 9197 | 1653 | 258 |
| 4 | Институт ядерной физики им. Г.И. Будкера СО РАН | www.inp.nsk.su | 23608 | 5850 | 2355 | 149 |
| 5 | Институт математики им. С.Л. Соболева СО РАН | www.math.nsc.ru | 4226 | 7233 | 1337 | 182 |
| 5 | Институт вычислительного моделирования СО РАН | icm.krasn.ru | 4915 | 5743 | 5751 | 474 |
| 7 | Государственная публичная научно-техническая библиотека СО РАН | www.spsl.nsc.ru | 5110 | 7653 | 418 | 136 |
| 8 | Институт систем информатики им. А.П. Ершова СО РАН | www.iis.nsk.su | 2352 | 13562 | 592 | 105 |
| 9 | Отделение ГПНТБ СО РАН | www.prometeus.nsc.ru | 4897 | 12370 | 241 | 94 |
| 10 | Институт автоматики и электрометрии СО РАН | www.iae.nsk.su | 2815 | 3983 | 3393 | 24 |
| 11 | Институт проблем освоения Севера СО РАН | www.ipdn.ru | 3637 | 9320 | 1541 | 57 |
| 12 | Новосибирский институт органической химии им. Н.Н. Ворожцова СО РАН | www.nioch.nsc.ru | 1789 | 4733 | 2384 | 16 |
| 13 | Институт катализа им. Г.К. Борескова СО РАН | www.catalysis.ru | 13441 | 178713 | 153 | 12 |
| 14 | Президиум СО РАН | www.sbras.nsc.ru | 5347 | 11827 | 1489 | 0 |
| 15 | Институт физики им. Л.В. Киренского СО РАН | www.kirensky.ru | 1424 | 3264 | 835 | 31 |

Таблица 5. Рейтинг сайтов организаций Общества Фраунгофера

| Рейтинг | Организация | Адрес сайта | V | S | R | Sc |
|---|---|---|---|---|---|---|
| 1 | Fraunhofer Headquarters | www.fraunhofer.de | 7209 | 16333 | 1247 | 624 |
| 2 | The Fraunhofer Institute for Systems and Innovation Research | www.isi.fraunhofer.de | 1548 | 3534 | 1449 | 464 |
| 3 | The Fraunhofer Institute for Open Communication Systems | www.fokus.fraunhofer.de | 1165 | 2456 | 588 | 298 |
| 4 | The Fraunhofer Institute for Manufacturing Engineering and Automation | www.ipa.fraunhofer.de | 1131 | 4565 | 488 | 143 |
| 5 | The Fraunhofer Institute for Industrial Mathematics | www.itwm.fraunhofer.de | 984 | 3017 | 865 | 212 |
| 6 | The Fraunhofer Institute for Solar Energy Systems | www.ise.fraunhofer.de | 2183 | 6495 | 543 | 243 |
| 7 | The Fraunhofer Institute for Industrial Engineering | www.iao.fraunhofer.de | 1287 | 2199 | 435 | 165 |
| 7 | The Fraunhofer Institute for Laser Technology | www.ilt.fraunhofer.de | 1072 | 2343 | 784 | 130 |
| 9 | The Fraunhofer Institute for Integrated Circuits | www.iis.fraunhofer.de | 4806 | 2309 | 669 | 521 |
| 10 | The Fraunhofer Institute for Information Center for Planning and Building | www.irb.fraunhofer.de | 2163 | 21078 | 125 | 95 |
| 11 | The Fraunhofer Institute for Factory Operation and Automation | www.iff.fraunhofer.de | 1319 | 2158 | 301 | 52 |
| 12 | The Fraunhofer Institute for Algorithms and Scientific Computing | www.scai.fraunhofer.de | 798 | 2116 | 490 | 206 |
| 13 | The Fraunhofer Institute for Building Physics | www.ibp.fraunhofer.de | 985 | 1519 | 695 | 83 |
| 14 | The Fraunhofer Institute for Intelligent Analysis and Information Systems | www.iais.fraunhofer.de | 938 | 2147 | 220 | 107 |
| 15 | The Fraunhofer Institute for Wind Energy and Energy System Technology | www.iwes.fraunhofer.de | 821 | 2947 | 391 | 63 |

Сравнительный анализ значений параметров **V**, **S**, **R** и **Sc** для сайтов научных организаций СО РАН и Общества Фраунгофера позволяет сделать следующие выводы.

Для 23 сайтов в СО РАН и для 18 сайтов в ОФ количество внешних ссылок на сайт превышает 1000. Таким образом, 24 % сайтов в СО РАН и 25 % сайтов в ОФ имеют достаточно много внешних ссылок. Более 100 веб-страниц содержат 84 % сайтов в СО РАН и 95 % сайтов в ОФ. Насыщенность сайтов как в СО РАН, так и в ОФ тоже очень близка: количество сайтов с параметром **R**, превышающим 100, для СО РАН равно 43,

а для ОФ — 48. По индексу цитирования Google Scholar сайты Общества Фраунгофера существенно опережают сайты СО РАН: количество сайтов с параметром **Sc**, превышающим 10, для СО РАН равно 36 (38 %), а для ОФ — 66 (92 %).

## Заключение

В работе рассмотрены веб-пространства трех академических сообществ – Сибирского отделения РАН (СО РАН), немецкого Общества Фраунгофера (ОФ) и сообщества Республики Сербия. Сайты входящих в них научных организаций ранжированы методами вебометрики. Определены численные и структурных характеристики соответствующих веб-графов, выявлены общие свойства и различия указанных академических веб-пространств.

Проведенный сравнительный анализ позволяет сформулировать рекомендации по дальнейшему формированию структуры веб-пространства СО РАН. По сравнению с Обществом Фраунгофера в веб-пространстве СО РАН в настоящее время много сайтов не достижимых друг из друга (из 21 сайта СО РАН невозможно по ссылкам попасть в остальные 74 сайта). Представляется необходимым увеличить число сайтов в сильно связной компоненте веб-графа СО РАН. Одним из средств достижения этой цели может быть более полное представление на сайтах научных организаций результатов исследований, в том числе по междисциплинарным проектам.

## Литература


1. Albert R., Barabási A.-L. Statistical mechanics of complex networks // Reviews of Modern Physics. 2002. V. 74, № 1. P. 47–97.
2. Almind T., Ingwersen P. Infometric analyses on the World Wide Web: Methodological approaches to “webometrics” // J. Document. 1997. Vol. 53, № 4. P. 404–426.
3. Thelwall M., Wilkinson D. Graph structure in three national academic webs: power laws with anomalies //Am. Soc. Inf. Sci. Technol. 2003. Vol. 54(8). P. 706–712.
4. Stuart D., Thelwall M., Harries G. UK academic web links and collaboration — an exploratory study // J. Inf. Sci. 2007. Vol. 33(2). P. 231–246.

5. Шокин Ю.И., Клименко О.А., Рычкова Е.В., Шабальников И.В. Рейтинг сайтов научных организаций СО РАН // Вычислительные технологии. 2008. Т. 13, № 3. С. 128–135.
6. Мазалов В.В., Печников А.А. О рейтинге официальных сайтов научных учреждений северо-запада России // Управление большими системами. Выпуск 24. М.: ИПУ РАН, 2009. С. 130–146.
7. Шокин Ю.И., Веснин А.Ю., Добрынин А.А., Клименко О.А., Рычкова Е.В., Петров И.С. Исследование научного веб-пространства Сибирского отделения Российской академии наук // Вычислительные технологии. 2012. Т. 17. № 6. С. 86–98.
8. Pechnikov A.A., Nwohiri A.M. Webometric analysis of Nigerian university websites // Webology. 2012. Vol. 9, № 1. (http://www.webology.org/2012/v9n1/a95.html)
9. Печников А.А. Применение вебометрических методов для исследования информационного веб-пространства научной организации (на примере Карельского научного центра РАН) // Труды КарНЦ РАН. No 1. Сер. Математическое моделирование и информационные технологии. Вып. 4. Петрозаводск: КарНЦ РАН, 2013. С. 86–95.
10. Веснин А.Ю., Константинова Е.В., Савин М.Ю. О сценариях присоединения новых сайтов к веб-пространству СО РАН // Вестник НГУ, серия: информационные технологии. 2013. Т. 11. № 4. С. 28–37.
11. Shokin Yu.I., Vesnin A.Yu., Dobrynin A.A., Klimenko O.A., Konstantinova E.V., Petrov I.S., Rychkova E.V. Investigation of the academic web space of the Republic of Serbia // Zbornik radova konferencije MIT 2013, Belgrad, Serbia, 2014. С. 601–607.
12. Шокин Ю.И., Веснин А.Ю., Добрынин А.А., Клименко О.А., Рычкова Е.В. Анализ веб-пространства академических сообществ методами вебометрики и теории графов // Информационные технологии. 2014. № 12. С. 31–40.
13. Рейтинг сайтов научных организаций СО РАН. http://www.ict.nsc.ru/ranking
14. Харари Ф. Теория графов. – М.: Мир, 1973.
15. Емеличев В.А., Мельников О.И., Сарванов В.И., Тышкевич Р.И. Лекции по теории графов. – М.: Наука, 1990.

16. Рейнгольд Э., Нивергельт Ю., Део Н. Комбинаторные алгоритмы. Теория и практика. – М.: Мир, 1980.
17. Связи научных организации Общества Фраунгофера. http://ousnano.sbras.ru/sitepage.php?PageID=2505 (дата доступа — 27.09.2013).
18. Информационная система «Организации и сотрудники СО РАН» http://www.sbras.ru/sbras/db/ (дата доступа — 27.09.2013).
19. Связи научных организаций СО РАН. http://ousnano.sbras.ru/sitepage.php?PageID=3008 (дата доступа — 27.09.2013).
20. Веб-граф институтов Сербии. http://ousnano.sbras.ru/sitepage.php?PageID=2506 (дата доступа — 07.04.2013).
21. Hage P., Harary F. Structural models in anthropology. Cambridge University Press: Cambridge, UK, 1983.
22. Watts D., Strogatz S. Collective dynamics of small world networks // Nature. 1998. Vol. 393. P. 440–442.
23. Opsahl T., Panzarasa P. Clustering in weighted networks // Social Networks. 1009. Vol. 31. P. 155–163.
24. Broder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A., Wiener J. Graph structure in the Web // Computer Networks. 2000. Vol. 33. № 1–6. P. 309–320.
25. Проект Ranking Web of World Research Centers. http://research.webometrics.info/
26. Поисковая система Яндекс. http://www.yandex.ru (дата доступа — 10.08.2014).
27. Поисковая система Google. http://www.google.ru (дата доступа — 10.08.2014).
28. Поисковая система Bing. http://www.bing.com (дата доступа — 10.08.2014).
29. Индекс цитирования каталога Яндекс. http://help.yandex.ru/catalogue/citation-index/tic-about.xml (дата доступа — 10.08.2014).
30. Система определения индекса цитирования в веб-пространстве Google Scholar. http://scholar.google.com (дата доступа — 10.08.2013).
31. Рейтинг сайтов научных организаций СО РАН. http://www.ict.nsc.ru/ranking/index.php?s_InfoID=15 (дата доступа — 10.08.2014).

32. Рейтинг сайтов институтов Общества Фраунгофера в Германии. http://www.ict.nsc.ru/sitepage.php?PageID=1000 (дата доступа — 27.09.2013).