УДК 004.93 **DOI** 10.15622/ia.2020.19.6.7

И.Ю. Еремеев, М.В. Татарка, Ф.Л. Шуваев, А.С. Цыганов

АНАЛИЗ МЕР ЦЕНТРАЛЬНОСТИ УЗЛОВ СЕТЕЙ НА ОСНОВЕ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ

Еремеев И.Ю., Татарка М.В., Шуваев Ф.Л., Цыганов А.С. **Анализ мер центральности узлов сетей на основе метода главных компонент.**

Аннотация. Анализ сетей разнообразной природы, которыми являются сети цитирования, а также социальные или информационно-коммуникационные сети, включает изучение топологических свойств, позволяющих оценивать взаимосвязи между узлами сети и различные характеристики, такие как плотность и диаметр сети, связанные подгруппы узлов и тому подобное. Для этого сеть представляется в виде графа – совокупности вершин и ребер между ними. Одной из важнейших задач анализа сетей является оценивание значимости узла (или в терминах теории графов – вершины). Для этого разработаны различные меры центральности, позволяющие оценить степень значимости вершин сетевого графа в структуре рассматриваемой сети.

Существующее многообразие мер центральности порождает проблему выбора той, которая наиболее полно описывает значимость центральность узла.

Актуальность работы обусловлена необходимостью анализа мер центральности для определения значимости вершин, что является одной из основных задач изучения сетей (графов) в практических приложениях.

Проведенное исследование позволило с использованием метода главных компонент среди известных мер центральности выявить коллинеарные меры, которые в дальнейшем можно исключать из рассмотрения. Это позволяет уменьшить вычислительную сложность расчетов, что особенно важно для сетей с большим числом узлов, и повысить достоверность интерпретации получаемых результатов при оценивании значимости узла в рамках анализируемой сети при решении практических задач.

Выявлены закономерности представления различных мер центральности в пространстве главных компонент, что позволяет классифицировать их с точки зрения близости образов узлов сети, формируемых в определяемом применяемыми мерами центральности пространстве.

Ключевые слова: метод главных компонент, мера центральности, граф, кластеризация, мера сходства

1. Введение. При анализе социальных сетей обеспечении информационной безопасности информационно-коммуникационных сетей (ИКС) и других приложений теории графов важную роль играет исследование мер центральности как отдельных узлов, так и сети в целом. На сегодняшний день известно около тридцати различных мер центральности, применяемых при анализе структуры и топологических свойств сетей [1-6], которые базируются на различных подходах к оценке степени важности узла сети. На рисунке 1 в качестве примера представлены три наиболее известные меры центральности с показателями значимости узлов.

Рассмотрим граф:

$$G = (V, E),$$

где V − вершины графа, а E – ребра (рис. 1а).

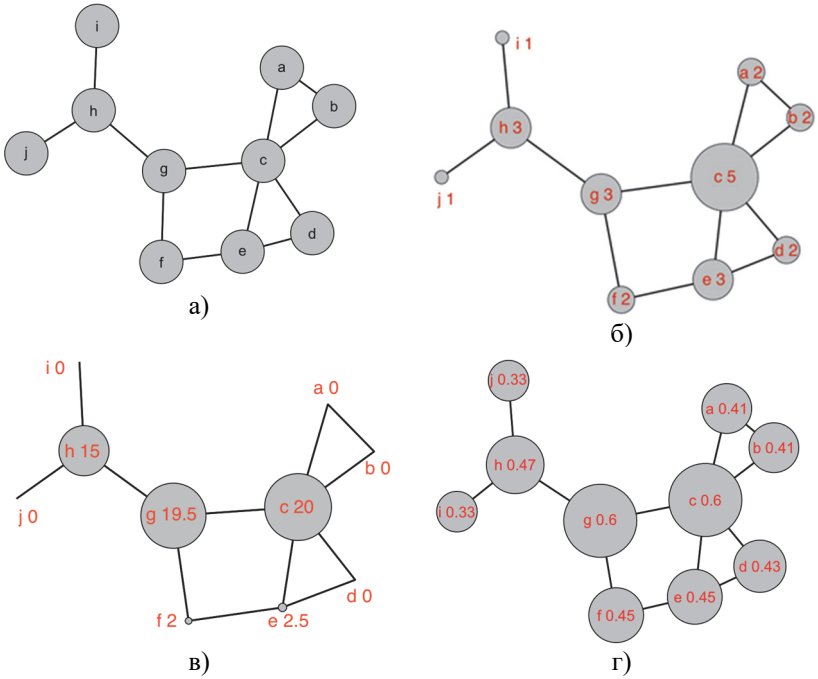

Рис. 1. Пример графов с различными мерами центральности вершин информационно-телекоммуникационных сетей: а) структура рассматриваемой ИКС; б) центральность по степени; в) центральность по посредничеству; г) центральность по близости

Будем полагать, что в рассматриваемом графе ребра являются ненаправленными и имеют одинаковый вес.

Центральность по степени определяется следующим образом (рис. 1б):

$$C_i = \deg(V_i),$$

где *deg* – степень *i*-го узла *V* графа *G* (число ребер узла).

Центральность по посредничеству (рис. 1в) характеризует, насколько важную роль вершина играет на пути «между» парами других вершин графа, в том смысле, что пути между другими вершинами должны проходить через рассматриваемую вершину. Вершина с наибольшим значением центральности по посредничеству является важной, так как позволяет от-

слеживать или контролировать поток информации в сети и может быть рассчитана в соответствии с выражением:

$$П_i = \sum_{v_j<v_k} g^*_{v_j v_k}(V_i)/g_{v_j v_k},$$

где $g^*_{v_j v_k}$ – общее число кратчайших путей из узла V_i и V_k; $g_{v_j v_k}(V_i)$ – число кратчайших путей, включающих вершину V_i.

Центральность по близости (рис. 1г) вершины графа определяется как величина, обратная сумме расстояний от узла V_i до всех остальных узлов:

$$Б_i = (\sum_{V_j=1}^{V} L(V_i, V_j))^{-1},$$

где $L(V_i, V_j)$ – расстояние между вершинами графа V_i и V_k.

Многообразие способов определения мер центральности позволяет получить огромное количество признаков, характеризующих анализируемые сети, что несомненно повышает требования к вычислительному ресурсу. В таких условиях возникает проблема выбора наиболее информативной меры или некоторого множества мер, позволяющих анализировать топологию исследуемых сетей.

Цель исследования – определить возможность сокращения признакового пространства, формируемого различными мерами центральности узлов анализируемых сетей. Данная тема исследована слабо, так как только в последние 2-3 года появились вычислительные мощности, достаточные для имитационного моделирования и исследования свойств сетей, которые имеют в своем составе тысячи вершин и связывающих их ребер [7-9].

В качестве информативного признака при исследовании выбран вклад мер центральности в формирование главных компонент при проведении компонентного анализа. Решается дополнительная задача, которая заключается в выявлении схожих мер центральности на основе метода главных компонент (МГК) и кластерного анализа. Выявленные группы мер центральности для разных типов сетей сравниваются между собой, что позволяет утверждать о взаимосвязи и однородности мер центральности.

Исследование основывается на базовых положениях теории графов и математической статистики. Совокупность узлов ИКС и связей между ними представлена в виде графа. При этом меры централь-

ности описывают потенциальную значимость вершин графа. Сравнительный анализ мер центральности проводится с использованием ряда реальных сетей, а также модельных сетей: на основе предпочтительного присоединения Барабаши – Альберта и модели малого мира Уоттса – Строгатца. Именно эти модели случайных графов наиболее полно описывают топологические свойства реальных сетей [10-13].

2. Модели сетей и меры центральности узлов. Для исследования использовались 8 моделируемых сетей: 1-4 – Уоттса – Строгатца, 5-8 – Барабаши – Альберта, и 8 реальных сетей.

2.1. Модель предпочтительного присоединения Барабаши – Альберта. Как показывает практика, новые вершины с большей вероятностью соединяются с вершинами, которые занимают выдающееся положение в сети, то есть имеют наивысшие показатели центральности по степени [11, 14]. Для анализа реальных ИКС разработаны свободно масштабируемые модели, самой распространенной среди которых является модель предпочтительного присоединения Барабаши – Альберта [14]. Формируется модель сети в виде графа по следующему принципу:

1. В начальный момент времени $t = 0$ есть V_t несвязных вершин.

2. На каждом шаге $(t = 1, 2, 3...)$ будем добавлять новую вершину с E_t ребрами.

3. Количество ребер, с которыми приходит в граф новая вершина, фиксировано, но соединяется она с уже существующей вершиной сети с вероятностью, пропорциональной степени этой вершины.

Максимально приближена к реальным сетям модификация модели Барабаши – Альберта с фиксированным параметром распределения вероятностей соединения вершин [10, 11, 14]. При таком варианте построения модели вводится кортеж распределения вероятностей образования вершин, в котором p_1 – вероятность изолированности вершины; p_2 – вероятность соединения вершины с одной вершиной; p_3 – вероятность соединения вершины с двумя вершинами. Рассмотренный кортеж заменяет этапы 2-3 построения модели Барабаши – Альберта в классическом варианте и позволяет регулировать соотношения между изолированными вершинами и вершинами с необходимым для моделирования количеством связей. На рисунке 2а изображен граф, построенный по модели Барабаши – Альберта без учета распределения вероятностей. Число вершин графа фиксировано и равно 500, а вновь появляющиеся вершины соединяются с вершинами с большей степенью. На рисунке 2б представлен граф, построенный по модели Барабаши – Альберта с уче-

том кортежа распределения вероятностей образования вершин, который имеет следующие значения: $p_1 = 0.25,\ \ p_2 = 0.5,\ \ p_3 = 0.25.$

а) б)

Рис. 2. Граф на основе модели Барабаши – Альберта: а) с равновероятно присоединенными вершинами; б) с заданными вероятностями присоединения вершин

На рисунке 2б видно, что граф содержит изолированные вершины и вершины, обладающие высокой степенью (увеличены в размерах). Таким образом, модель Барабаши – Альберта возможна в двух вариантах: одно- и двухпараметрическом. Параметрами являются число вершин V и вероятность присоединения фиксированного числа вершин p на каждом шаге моделирования [10, 11].

2.2. Модель малого мира Уоттса – Строгатца. Предложена в 1998 году американскими учеными Д. Уоттсом и С. Строгатцем [15-18]. Модель представляет собой одномерную регулярную решетку, состоящую из V вершин, причем каждая из них соединена только с k-ближайшими соседями, и на нее наложены периодические граничные условия (вероятность p соединения с любыми другими вершинами, кроме соседних, равна нулю), то есть при $p = 0$ решетка свернута в кольцо. После этого каждая связь с заданной вероятностью p перебрасывается на другую случайно выбранную вершину. При $p = 0$ получаем граф с исходной решеткой. На рисунке 3 представлены графы на основе модели малого мира Уоттса – Строгатца. Эти графы имеют четыре различные вероятности переключения ребер, что приводит к различным топологиям графа.

Таким образом, модель малого мира является трехпараметрической, управляемой тремя параметрами: количеством вершин V, количеством соседних узлов для каждой вершины k, вероятностью соединения с другими вершинами p.

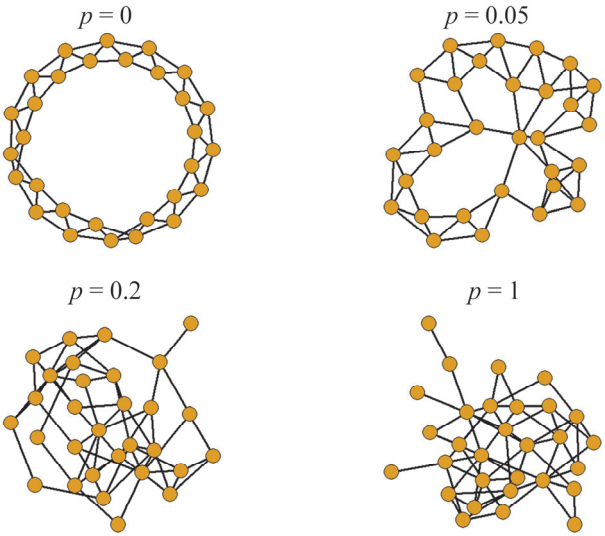

Рис. 3. Граф на основе модели малого мира Уоттса – Строгатца с изменяемой вероятностью переключения ребер

2.3. Используемые реальные сети. Реальные сети выбраны из пакета «нетворкдата» (англ. «networkdata») – самого крупного набора сетевых данных для языка статистического моделирования «R». Характеристики сетей представлены в таблице 1. Сети преобразованы к ненаправленным.

Выбранные реальные данные отражают различные стороны человеческой деятельности, а именно: сети 9-10 ИКС в датацентрах одной из коммерческих организаций; 11 – сеть межбелковых взаимодействий в человеческом организме; 12 – сеть взаимодействий между участниками выставки в Дублине в 2009 году; 13 – сеть дружеских отношений между студентами Австралийского национального университета; 14 – данные фиксируют случаи с местных пресс-релизов, выпущенных сенаторами США; сети 15-16 – взаимодействие между персонажами в произведениях Уильяма Шекспира.

Таблица 1. Характеристики исследуемых сетей

Номер сети\ Характеристика	Число вершин	Число ребер	Дополнительные параметры
1	100	100	$p = 0,01;\ k = 2$
2	100	100	$p = 0,05;\ k = 2$
3	100	300	$p = 0,1;\ k = 3$
4	100	400	$p = 0,2;\ k = 4$
5	100	143	$p = \{0;0,5;0,5\}$
6	100	424	$p = \{0;0,25;0,15;0,1;0,15;0,1;0,2;0,5\}$
7	100	265	$p = \{0;0,25;0,25;0,25;0,25\}$
8	100	99	Без дополнительных параметров
9	27	69	Без дополнительных параметров
10	26	137	Без дополнительных параметров
11	212	244	Без дополнительных параметров
12	410	17298	Без дополнительных параметров
13	216	2672	Без дополнительных параметров
14	92	477	Без дополнительных параметров
15	47	228	Без дополнительных параметров
16	52	314	Без дополнительных параметров

2.4. Исследуемые меры центральности. Наиболее широко распространены около 30 мер центральности [19-21], которые активно продвигаются своими разработчиками. Однако в настоящее время отсутствует четкое понимание алгоритмов выбора и применения мер при различных условиях для определения величины центральности узла графа. Для более наглядного графического представления проводимых в исследовании экспериментов мерам центральности присвоены сокращения (табл. 2) [20, 21].

Таблица 2. Исследуемые меры центральности

№ п/п	Мера центральности	Сокращение	№ п/п	Мера центральности	Сокращение
1	По степени	С	13	Центральность подграфа	ЦП
2	Взвешенная степень вершины	ВСВ	14	Среднее расстояние узла	СР
3	По посредничеству	П	15	Барицентр	БЦ
4	По близости	Б	16	Близость по Фриману	БФ
5	Собственный вектор	СВ	17	Близость по Латору	БЛ
6	Ранг страницы	РС	18	Близость остатков	БО
7	Эксцентриситет	ЭКС	19	Межкликовая связь	МКС
8	Авторитетность Кляйнберга	АК	20	Плотность максимальной компоненты окрестности узла	ПМКО
9	Концентрация Кляйнберга	КК	21	Линейная центральность	ЛЦ
10	Геодезическая центральность	ГЦ	22	Марковская центральность	МЦ
11	Центральность по вектору Лапласа	ЦЛ	23	Радиальная центральность	РЦ
12	Центральность «рычага»	ЦР	–	–	–

В ходе проведенного исследования были выбраны 23 меры центральности, которые представлены в библиотеке «centiserve» языка программирования «R» [19-21]. Предпочтение отдавалось мерам центральности, не требующим дополнительного параметра для настройки.

3. Основные этапы сравнительного анализа. На рисунке 4 представлена структурная схема сравнительного анализа мер центральности узлов графа на основе метода главных компонент. Он состоит из трех основных этапов, каждый из которых разделяется на подэтапы.

3.1. Этап подготовки данных. При решении задач машинного обучения и анализа данных около 30-40 % времени уходит на подготовку данных. Некорректно подготовленные данные могут существенно исказить результаты эксперимента.

1.1. Предобработка данных. Включает ряд преобразований.

– извлекается гигантская компонента графа, то есть извлекается максимальный связный подграф. Данное действие необходимо для избавления от висячих вершин и преобразования графа к связному;

– удаляются петли графа, то есть ребра, инцидентные одной и той же вершине;

– полученный граф преобразуется к неориентированному.

Рассмотренные преобразования не снижают достоверности дальнейшего анализа в силу того, что большинство мер центральности рассчитываются для связных и неориентированных графов.

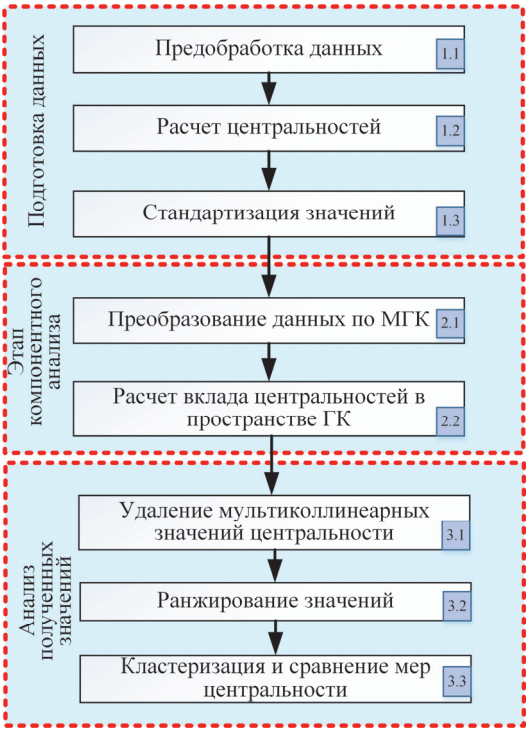

Рис. 4. Структурная схема сравнительного анализа меры центральности узлов сетей на основе метода главных компонент

1.2. Расчет мер центральности осуществляется для каждой вершины каждого из рассматриваемых графов (табл. 1). Полученные значения y_{ij} помещаются в список данных. При этом каждая ячейка списка соответствует одной исследуемой сети и представляет собой матрицу рассчитанных значений N мер центральности для n вершин сетевого графа:

$$Y_{[N,n]} = \left\| y_{ij} \right\|_N^n .$$

1.3. Стандартизация значений. Результаты вычислений центральности имеют различную размерность, поэтому выполняется их

нормировка по величине среднеквадратического отклонения значений центральности узлов для каждой меры.

3.2. Этап компонентного анализа. Компонентный анализ в рамках метода главных компонент (МГК) представляет собой совокупность статистических приемов обработки данных, которые позволяют сконцентрировать содержащуюся в исходном массиве данных информацию за счет перехода к меньшему числу наиболее информативных факторов – главных компонент (ГК).

2.1. Преобразование данных по МГК. Исходными данными являются стандартизированные значения y_{ij}. Каждую строку матрицы можно представить как реализацию n-мерного случайного вектора:

$$\hat{Y}_{\langle n\rangle} = \langle \hat{y}_1, \hat{y}_2, ..., \hat{y}_n \rangle.$$

Свойства данного случайного вектора с достаточной для практики точностью описываются вектором математических ожиданий:

$$\overline{Y}_{\langle n\rangle} = \langle \overline{y}_1, \overline{y}_2, ..., \overline{y}_n \rangle$$

и корреляционной матрицей, содержащей линейные коэффициенты корреляции Пирсона:

$$K_{[n]} = \left\| K_{j1,j2} \right\|_n^n, \text{ где } K_{j_1,j_2} = \frac{\sum_{i=1}^{N} (\hat{y}_{j1,i} - \overline{y}_{j1})(\hat{y}_{j2,i} - \overline{y}_{j2})}{\sqrt{\sum_{i=1}^{N} (\hat{y}_{j1,i} - \overline{y}_{j1})^2 \sum_{i=1}^{N} (\hat{y}_{j2,i} - \overline{y}_{j2})^2}}.$$

Метод ГК основывается на предположении, что любой j-й признак может быть представлен в виде линейной комбинации ГК f_i:

$$y_j = a_{1,j} f_1 + a_{2,j} f_2 + ... + a_{n,j} f_n, \ \left[j = \overline{1(1)n} \right],$$

где $f_1, f_2, ..., f_n$ – главные компоненты; $a_{m,j}$ – вес m-й ГК в j-м признаке.

Главные компоненты рассчитываются таким образом, чтобы первая из них давала максимально возможный вклад в суммарную дисперсию наблюдений, вторая – максимальный вклад в дисперсию, оставшуюся в суммарной дисперсии за вычетом первой главной компоненты и так далее. Таким образом, задача анализа главных

компонент сводится к тому, чтобы найти такое линейное ортогональное преобразование n наблюдаемых признаков, которое позволит получить совокупность n некоррелированных нормированных переменных $f_i, [i = 1(1)n]$, дисперсии σ^2 которых обладают следующим свойством:

$$\sigma_1^2 \geq \sigma_2^2 \geq ... \geq \sigma_n^2.$$

Такое преобразование эквивалентно преобразованию исходной корреляционной матрицы $K_{[n]}$ к матрице вида:

$$K_{[n]}^T K_{[n]} = \begin{vmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \sigma_n^2 \end{vmatrix}.$$

По дисперсии, соответствующей каждой ГК, можно оценить вклад этой компоненты в формирование общей дисперсии совокупности результатов расчета мер центральности и тем самым выделить существенные компоненты. При этом возможны следующие варианты использования результатов вычисления ГК:

– расчет вклада каждого наблюдаемого признака в ГК и их ранжирование по степени вклада;

– выявление мультиколлинеарных признаков, то есть имеющих тесную корреляционную взаимосвязь;

– разделение наблюдаемых признаков по группам (кластеризация).

2.2. Для каждой из исследуемых сетей осуществляется расчет вклада различных мер центральности в пространстве ГК. Для расчета удельного вклада каждой меры центральности используется следующее выражение [22]:

$$Con_i = \frac{f_1(y_i) f_2(y_i)}{\sum_{k=1}^{I} f_1(y_k) f_2(y_k)}, \ [i = 1(1)N],$$

где $f_1(y_i)$ и $f_2(y_i)$ – значение вклада i-й меры в ГК1 и ГК2.

Таким образом, значение удельного вклада *Con* (от англ. «contribution») является произведением вклада в ГК1 и ГК2 *i*-й меры центральности, нормированной на сумму вкладов остальных мер в ГК1 и ГК2.

3.3. Этап анализа полученных значений. Заключительный этап анализа. На данном этапе полученные результаты интерпретируются.

3.1. Выделение коллинеарных значений центральности. Рассчитанные значения вклада центральностей записываются в матрицу. Чтобы минимизировать ошибку на последующих этапах метода, необходимо сократить признаковое пространство, то есть выявить и удалить дублирующие друг друга значения *Con* различных мер центральности.

3.2. Ранжирование полученных значений. Массив со значениями вкладов мер центральности (с учетом предыдущего преобразования) ранжируется для каждой сети.

3.3. Кластеризация и сравнение мер центральности. На данном этапе производится кластеризация рассчитанных значений мер центральности в пространстве ГК f_1 и f_2, в ходе которой схожие меры разбиваются на группы методом иерархической кластеризации. Далее сравнивается состав кластеров мер центральности для всех исследуемых сетей. В качестве меры схожести результатов кластеризации используется коэффициент Фоулкса – Мэллова [23].

4. Апробация метода. Полученные в результате этапа 2 значения вклада ГК представлены в таблице 3. Результаты анализа полученных значений на подэтапе 3.1 – 3.2 представлены в таблице 4.

С данными, полученными в результате имитационного моделирования, и с реальными данными производились преобразования в соответствии с этапами сравнительного анализа, представленными на рисунке 4. На подэтапе 1.1 в результате предобработки каждая из 16 сетей (табл. 1) приведена к виду связного неориентированного графа без петель. На подэтапе 1.2 для каждого графа рассчитаны 23 меры центральности вершин (табл. 2). Полученные данные стандартизировались, после чего подавались на второй этап анализа.

На втором этапе выполнен расчет вклада каждой меры центральности в первую и вторую главные компоненты (табл. 3).

Таблица 3. Расчет вклада мер центральностей в первую и вторую главные компоненты

	БЛ	ЛЦ	БЦ	БФ	Б	МЦ	РЦ	СР	БО	ЖС	СВ	АК	КК	П	ЦЛ	ВСВ	С	ГМК	РС	ЦР	МКС
1	7,95	7,67	7,67	7,67	7,67	7,61	7,13	7,13	6,83	6,78	5,49	5,49	5,49	4,08	1,47	0,99	0,99	0,86	0,67	0,18	0,08
2	7,32	6,48	6,48	6,48	6,48	7,39	6,25	6,25	7,33	3,85	3,29	3,29	3,29	4,71	4,75	4,01	4,01	0,27	3,15	1,80	1,07
3	6,14	5,54	5,54	5,54	5,54	6,48	5,42	5,42	5,96	1,33	4,26	4,26	4,26	4,94	5,95	5,53	5,53	0,02	5,11	3,53	0,95
4	5,53	5,17	5,17	5,17	5,17	5,61	5,14	5,14	5,40	0,67	5,12	5,12	5,12	4,55	5,51	5,34	5,34	0,13	5,18	4,18	1,75
5	5,30	5,03	5,03	5,03	5,03	5,12	4,26	4,26	5,26	3,22	5,33	5,33	5,33	4,40	5,05	4,86	4,86	0,99	4,52	2,10	4,90
6	5,12	5,06	5,06	5,06	5,06	5,14	4,59	4,59	5,00	2,85	5,10	5,10	5,10	3,90	4,98	5,13	5,13	0,01	5,04	4,29	4,28
7	5,24	5,14	5,14	5,14	5,14	5,13	4,40	4,40	5,12	2,72	5,29	5,29	5,29	4,11	4,94	5,07	5,07	0,05	4,92	3,21	4,80
8	5,75	4,95	4,95	4,95	4,95	5,73	3,98	3,98	5,74	3,54	4,63	3,09	3,09	4,62	5,64	4,83	4,83	4,83	4,62	1,21	4,83
9	5,41	4,82	4,82	4,68	4,68	5,78	3,55	3,55	4,89	1,24	4,12	4,12	4,12	5,54	5,85	5,14	5,82	0,90	5,23	4,30	5,51
10	5,41	5,36	5,36	5,47	5,47	6,08	4,98	4,98	5,30	1,73	2,18	2,18	2,18	3,28	6,07	4,20	6,05	4,86	4,12	4,99	4,24
11	5,76	5,12	5,12	5,12	5,12	5,44	4,42	4,42	5,77	3,74	4,65	4,65	4,65	4,40	5,10	4,27	4,27	3,61	3,36	1,97	4,26
12	6,83	6,18	6,18	6,18	6,18	6,48	5,89	5,89	6,70	3,91	2,88	2,88	2,88	1,55	6,01	6,01	6,01	0,02	4,57	3,28	1,27
13	4,77	5,34	5,34	5,32	5,32	6,56	5,02	5,02	4,05	1,63	4,18	4,18	4,18	2,95	6,46	5,07	6,53	0,98	4,87	4,81	2,23
14	5,04	4,99	4,99	4,99	4,99	5,03	4,42	4,42	4,91	2,66	5,03	5,03	5,03	3,26	4,85	5,04	5,04	3,64	4,92	3,89	3,47
15	5,56	4,80	4,80	4,39	4,39	6,14	4,08	4,08	5,15	2,91	4,47	4,47	4,47	4,44	6,22	4,37	5,83	2,99	4,28	3,16	3,18
16	5,42	5,21	5,21	5,13	5,13	5,79	4,63	4,63	5,27	1,05	3,39	3,39	3,39	3,27	5,61	4,95	5,84	3,74	4,99	5,08	3,70

Таблица 4. Ранжированные значения вклада мер центральности для анализируемых сетей

Ц /№ сети	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
МЦ	1	1	1	1	2	3	1	9	4	1	2	1	2	3	2	2
БЛ	2	4	2	5	7	6	7	8	9	6	4	2	12	6	5	4
БО	3	2	6	7	8	10	11	7	10	8	3	5	13	10	9	7
ЛЦ	4	8	9	9	10	9	10	13	11	7	11	8	8	8	11	8
Б	5	7	8	8	11	8	9	14	13	5	12	7	9	9	12	10
РЦ	6	12	10	11	14	11	14	15	15	10	15	9	11	12	15	12
КК	7	10	13	10	3	5	2	11	14	16	14	13	10	5	10	16
П	8	9	11	14	13	14	13	12	6	15	9	15	14	16	8	15
ЭКС	9	13	15	16	15	16	16	16	16	17	16	12	16	17	17	17
ЦЛ	10	3	3	2	1	7	5	1	1	2	1	6	3	7	1	3
ВСВ	11	6	5	3	5	2	3	4	8	13	6	3	4	2	6	11
С	12	5	4	4	4	1	4	3	2	3	5	4	1	1	3	1
РС	13	11	7	6	9	4	6	10	7	14	13	10	5	4	7	9
ЦР	14	15	12	12	16	13	15	17	12	9	17	11	6	13	13	6
ЦП	15	14	14	13	12	12	12	2	3	4	8	14	7	11	4	5
МКС	16	16	16	15	6	15	8	5	5	12	7	16	15	15	14	13
ПМКО	17	17	17	17	17	17	17	6	17	11	10	17	17	14	16	14

На подэтапе 3.1 для поиска коллинеарных значений проанализирован расчет вклада центральностей (табл. 3). Выявлено 9 коллинеарных мер центральности, заштрихованных в таблице 3. Коллинеарные меры центральности распределены по следующим группам:

– первая группа: ЛЦ || БЦ || БФ || Б;
– вторая группа: РЦ || СР;
– третья группа: СВ || АК || КК.

Меры ВСВ и С коллинеарны частично, поэтому исключать их из дальнейших расчетов не будем.

После удаления коллинеарных мер БЦ, БФ, Б, СР, СВ, АК, оставшиеся меры центральности узлов ранжируются в зависимости от вклада *Con* для каждой сети. Результаты такого преобразования представлены в таблице 4. Среди мер центральности видны явные аутсайдеры – это меры центральности ПМКО, МКС, ЭКС. Первую позицию, в большинстве сетей занимает мера марковской центральности (МЦ).

МЦ – это мера, основанная на концепции случайного обхода графа. В ней используется среднее время первого прохождения от каждой вершины до каждой другой вершины. Оно показывает насколько тесно каждая вершина связана с другой. Среднее время первого прохода от вершины a к вершине b – это среднее число шагов,

получаемые в процессе случайного блуждания. Случайные блуждания с большей вероятностью быстрее достигают вершин, занимающих главенствующее положение в сети.

В результате изучения публикаций по данной предметной области установлено, что МЦ незаслуженно обделена вниманием исследователей. Так, в русскоязычном сегменте сети Интернета не найдено ни одного научного труда, посвященного этой мере центральности. В англоязычном – статьи, датированные 2003 годом [24].

На рисунках 5 и 6 представлены корреляционные окружности для некоторых из исследуемых сетей. Корреляционная окружность – это способ визуализации результатов анализа данных методом главных компонент. Окружность имеет единичный радиус, что соответствует суммарному вкладу первой и второй компонент. Чем ближе к ней значение координат вектора, характеризующего ту или иную меру центральности, тем выше его вклад в указанные компоненты. Схожие по степени вклада в ГК1 и ГК2 меры центральности расположены рядом.

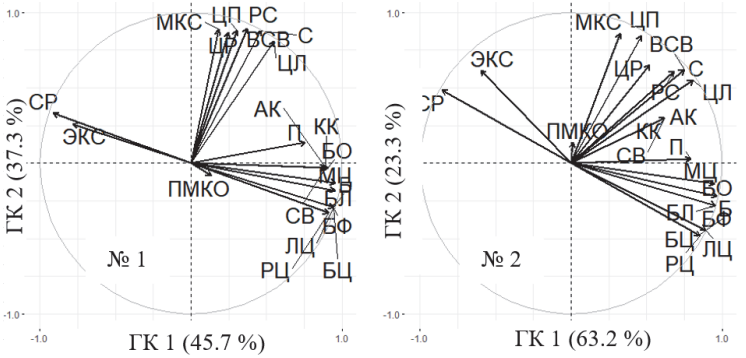

Рис. 5. Корреляционные окружности для моделей сетей № 1-2

На рисунке 5 корреляционные окружности сверху: для моделей сетей 1 и 2, построенных на базе модели Уоттса – Строгатца (слева) и на базе модели Барабаши – Альберта (справа). На рисунке 5 видно, что для каждой сети обособлена мера СР и ЭКС. Остальные меры имеют склонность к группированию и сильно коррелированы между собой.

Представленные на рисунке 6 корреляционные окружности относятся реальным сетям 9 (слева) и 10 (справа). Обособленность мер ЭКС и ПМКО сохраняется и для реальных сетей. Остальные меры так же как в смоделированных сетях склонны образовывать группы и имеют положительную высокую корреляцию между собой.

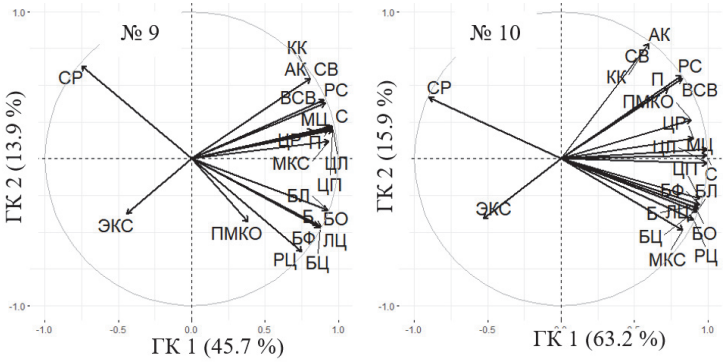

Рис. 6. Корреляционные окружности для реальных сетей № 9-10

Закономерности группировки данных позволяют применить методы кластерного анализа для оценивания близости мер центральности. Результатом иерархической кластеризации мер центральности в пространстве ГК является дендрограмма или дерево, построенное по матрице сходства, которая определяет расстояние между парами кластеров. Самыми распространенными методами расчета расстояния являются метод одиночной связи, метод полной связи, метод средней связи, центроидный метод и метод Уорда. Для выбора метода расчета расстояния между кластерами воспользуемся коэффициентом попарной корреляции Пирсона строк матрицы сходства и матрицы кофенетического расстояния результатов кластеризации, выполненной с использованием различных методов расчета расстояния. Результаты расчетов приведены в таблице 5.

Таблица 5. Значения коэффициента корреляции для различных методов расчета расстояния

Вид расстояния	Уорда	Одиночной связи	Полной связи	Средней связи	Центроидный
Корреляция	0,95	0,867	0,789	0,829	0,851

Из таблицы 5 следует, что наилучшим является расстояние Уорда, так как корреляция между матрицей расстояний Уорда и матрицей кофенетического расстояния наибольшая. Следовательно, расстояние Уорда и будем применять при построении дендрограмм. Результаты иерархической кластеризации представлены в виде дендрограмм на рисунках 7 и 8.

На рисунке 7 видно, что группы мер центральности, полученные в результате кластеризации, в основном схожи. Из общей картины выделяются кластеры ЭКС и ПМКО, число кластеров для всех сетей

равно четырем. Таким образом, меры центральности делятся на 4 кластера, из которых можно выделить два крупных и два малых. Малые кластеры имеют различный состав, что позволяет их отнести к случайным выбросам. Для выявления различий между кластерами, образованными мерами центральности, рассчитаем их попарное сходство. Для этого будем использовать коэффициент Фоулкса – Мэллова [23].

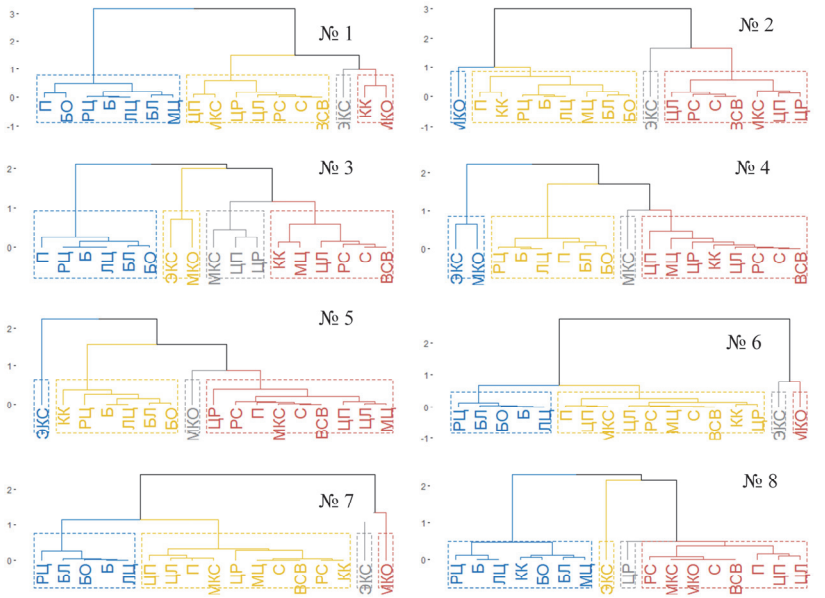

Рис. 7. Результаты кластеризации мер центральности для исследуемых сетей, построенных на базе моделей (№ 1-8)

Коэффициент Фоулкса – Мэллова – это мера сходства между двумя результатами кластеризации. Рассмотрим, как рассчитывается коэффициент Фоулкса – Мэллова. При сравнении результатов кластеризации при условии одинакового количества кластеров возможны 4 варианта положения объектов кластеризации (точек):

1. *YY* – количество объединенных в одинаковые кластера точек при кластеризации A_1 и A_2.
2. *YN* – количество точек присутствующих в одном кластере при кластеризации A_1, но не A_2.
3. *NY* – количество точек присутствующих в одном кластере при кластеризации A_2, но не A_1.
4. NN – количество точек не совпадающих ни в A_1 ни в A_2.

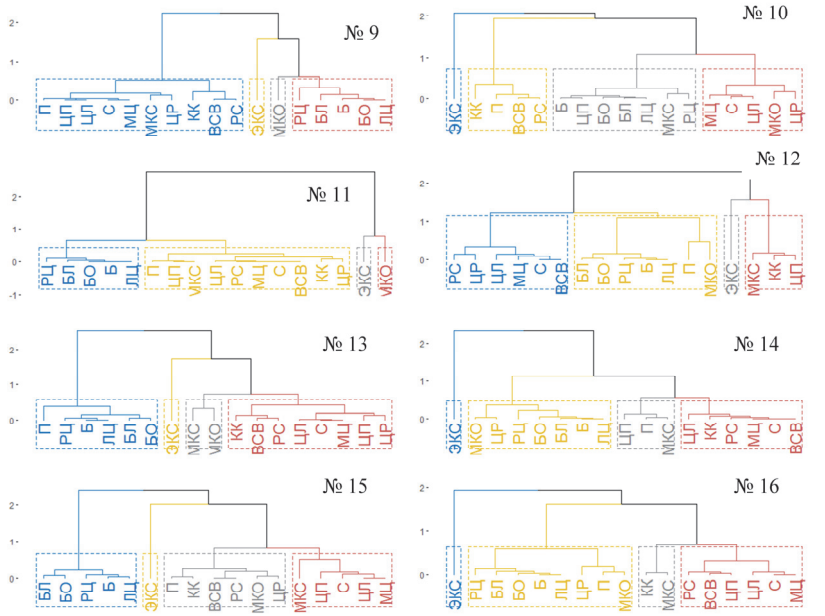

Рис. 8. Результаты кластеризации мер центральности для исследуемых реальных сетей (№ 9-16)

Коэффициент Фолукса – Мэллова рассчитывается согласно выражению:

$$C = \frac{YY}{\sqrt{(YY + YN) \times (YY + NY)}}.$$

Коэффициент изменяется в пределах $[-1;1]$. Чем его значение выше, тем более похожи друг на друга результаты кластеризации.

Результаты расчета коэффициента Фоулкса – Мэллова представлены в виде матрицы попарных сравнений на рисунке 9.

Таким образом, можно утверждать, что рассматриваемые меры центральности узлов сгруппированы примерно одинаково для всех исследуемых сетей, что говорит об устойчивости полученных результатов. По матрице попарных сравнений можно сделать вывод, что группировка мер центральности имеет схожесть для всех исследуемых 16 сетей. Значения коэффициента Фоулкса – Мэллова не опускаются ниже 0,41, а максимальные значения в ряде случаев достигают 1.

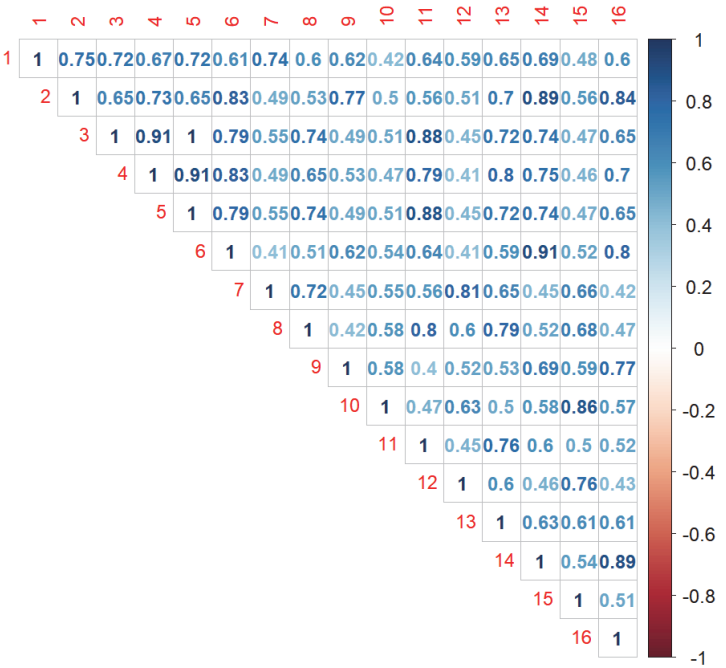

Рис. 9. Матрица попарных сравнений меры сходства кластеров на основе коэффициента Фоулкса – Мэллова

Рассчитаем статистическую значимость положения меры в ранжированном списке ее вклада в ГК. На рисунке 10 в виде столбцовой диаграммы показана зависимость статистической значимости для мер центральности от значения вклада в первые три ГК. Статистическая значимость рассчитывалась в соответствии с критерием согласия Пирсона. Выбор только трех ГК обусловлен тем, что в конкретном исследовании они определяют 95% обобщенной дисперсии данных.

Как видно на рисунке 10, наибольший вклад при оценке важности узла сети вносят меры ЦП, С, МЦ. Значения по вертикальной оси пропорциональны наблюдаемым частотам встречаемости меры центральности на соответствующей позиции значимости по всем исследуемым графам.

Цветом обозначена значимость отклонения ожидаемых и наблюдаемых частот в этой ячейке, если значения нормализованных стандартизированных остатков больше 0.025, можно считать, что в этой ячейке зафиксированы статистически значимые отклонения, что говорит о степени важности меры центральности в контексте всех исследованных сетей.

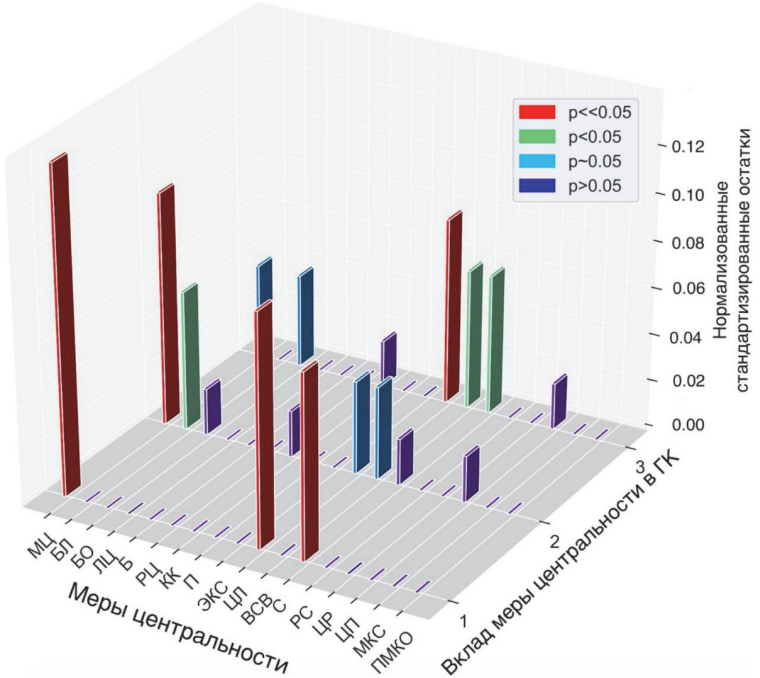

Рис. 10. Значения вклада мер центральности в первые три компоненты по всем исследуемым графам сетей

При этом под ожидаемыми частотами принимаются вычисленные по критерию Хи-квадрат равновероятные частоты, иначе говоря, отклонение расстояния Пирсона в значимых пределах позволяет сказать о статистической значимости положения меры центральности в ранжированном списке ее вклада в ГК в сравнении с равновероятным распределением.

Таким образом, на основании проведенного сравнительного анализа статистически обоснована значимость мер центральности, что позволяет в дальнейшем начинать анализ структуры ИКС с наиболее значимых мер. Кроме того, обоснована возможность сокращения признакового пространства, формируемого различными мерами центральности узлов при анализе ИКС различной природы.

5. Заключение. Полученные результаты составляют определенный вклад в развитие приложений теории графов в задачах исследования сетей разнообразной природы и обосновывают возможность сокращения признакового пространства, формируемого различными мерами центральности узлов анализируемых сетей.

Использован набор данных о модельных и реальных информационно-коммуникационных сетях различной природы, для которых меры центральности узлов имеют достаточно высокий уровень статистической значимости, что соответствует коэффициенту кластеризации на уровне не менее 0,7.

Анализ 23 мер центральности узлов сети методом главных компонент показал, что 9 исследуемых мер центральности являются коллинеарными. Коллинеарные меры центральности распределены по трем группам. Первая группа включает линейную центральность, барицентральность, центральность по Фриману и по близости. Вторая группа – центральность по среднему расстоянию и центральность «рычага». Третья группа – центральность по собственному вектору, центральность по авторитетности Кляйнберга и по концентрации Кляйнберга.

В результате ранжирования вклада мер центральности в значения главных компонент для каждой сети выявлено, что меры центральности по плотности максимальной компоненты окрестности узла, межкликовой связи и эксцентриситету являются малозначимыми, а первую позицию в большинстве сетей занимает марковская центральность.

Обоснованное применение метода Уорда для расчета расстояния между кластерами, формируемыми различными мерами центральности в пространстве главных компонент позволило, осуществив кластеризацию, сделать вывод о группировании мер центральности по четырем кластерам. По матрице попарных сравнений результатов кластеризации, полученной с использованием коэффициента Фоулкса – Мэллова, можно сделать вывод о том, что группировка мер центральности имеет схожесть для всех исследуемых 16 сетей, что говорит об устойчивости полученных результатов.

Наибольший вклад в результат оценивания важности узла сети вносят меры центральности по степени, марковской центральности и центральности подграфа.

Для дальнейших исследований представляют интерес меры центральности по среднему расстоянию узла и по эксцентриситету, так как они расположены обособленно в пространстве главных компонент относительно остальных мер центральности, а также меры центральности по степени, марковской центральности и центральности подграфа, как наиболее значимые при оценке важности узла сети.

Литература

1. *Bonchi F., De Francisci G., Riondato M.* Centrality Measures on Big Graphs: Exact, Approximated, and Distributed Algorithms // Proceedings of the 25th International Conference Companion on World Wide Web. 2016. pp. 1017–1020.

2. *Щербакова Н.Г.* Меры центральности в сетях // Проблемы информатики. 2015. № 1. С. 18–30.
3. *Бередихин С.В., Ляпунов В.М., Щербакова Н.Г.* Мера важности научной периодики – «Центральность по посредничеству» // Проблемы информатики. 2014. № 3. С. 53–63.
4. *Юдина М.Н.* Узлы в социальных сетях: меры центральности и роль в сетевых процессах // Омский научный вестник. 2016. № 4. С. 161–165.
5. *Brandes U., Borgatti S., Freeman L.* Maintaining the duality of closeness and betweenness centrality // Social Networks. 2016. vol. 44. pp. 153–159.
6. *Minoo A. et al.* A Systematic Survey of Centrality Measures for Protein-Protein Interaction Networks // BMC Systems Biology. 2018. vol. 12. no. 1. pp. 80.
7. *Chen P-Y., Choudhury S., Hero A.,* Multi-centrality graph spectral decompositions and their application to cyber intrusion detection // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. pp. 4553–4557.
8. *Lu B., Sun H., Harris P., Xu M.* Shp2graph: Tools to Convert a Spatial Network into an Igraph Graph in R // ISPRS Int. J. Geo-Inf. 2018. vol. 7. pp. 293.
9. *Csardi G, Nepusz T.* The IGRAPH software package for complex network research // InterJournal, Complex Systems. 1695. 2006. vol. 1695. no. 5. pp. 1–9.
10. *Шуваев Ф.Л., Татарка М.В.* Анализ динамики мер центральности математических моделей случайных графов // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 20. № 2. С. 249–256.
11. *Шуваев Ф.Л., Татарка М.В.* Анализ математических моделей случайных графов, применяемых в имитационном моделировании информационно-коммуникационных сетей // Вестник Санкт-Петербургского университета ГПС МЧС России. 2020. № 2. С. 67–77.
12. *Van Mieghem P., Ge X., Schumm P., Trajanovski S., Wang H.* Spectral graph analysis of modularity and assortativity // Phys. Rev. 2010. vol. 82. no. 5. P. 056113.
13. *Barzel B., Biham O.* Quantifying the connectivity of a network: the network correlation function method // Phys. Rev. 2009. vol. 80. pp. 046104.
14. *Barabasi A.* Network Science // Cambridge university press. 2016. 453 p.
15. *Watts D., Strogatz H.* Collective dynamics of «Small-world» networks // Nature. 1998. vol. 393. pp. 440–442.
16. *Hartmann A., Mézard M.* Distribution of diameters for Erdős-Rényi random graphs // Phys. Rev. 2018. vol. 97. no. 3. pp. 032128.
17. *Le C., Levina E., Vershynin R.* Concentration and regularization of random graphs // Random Structures&Algorithms. 2017. vol. 51. no. 3. pp. 538–561.
18. *Gibson H., Vickers P.* Using adjacency matrices to lay out larger small-world networks // Applied soft computing. 2016. vol. 42. pp. 80–92.
19. *Jalili M. et al.* CentiServer: A Comprehensive Resource, Web-Based Application and R Package for Centrality Analysis // PLoS ONE. 2015. vol. 10. no. 11. pp. 0143111.
20. *Oldham, S. et al.* Consistency and differences between centrality measures across distinct classes of networks // PLoS ONE. 2019. vol. 14. no. 7. pp. 0220061.
21. *Bloch F., Jackson M., Tebaldi P.* Centrality measures in networks // SSRN. 2016. 42 p.
22. *Lê S., Josse J., Husson F.* FactoMineR: A Package for Multivariate Analysis // Journal of Statistical Software. 2008. vol. 25. no. 1. pp. 1–18.
23. *Depaolini M., Ciucci D., Calegari S., Dominoni M.* External Indices for Rough Clustering // Lecture Notes in Computer Science. 2018. vol. 11103. pp. 378–391.
24. *White S., Smyth P.* Algorithms for estimating relative importance in networks // Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003. pp. 266–275.

Еремеев Игорь Юрьевич – д-р техн. наук, профессор, профессор, Военно-космическая академия имени А.Ф. Можайского (ВКА им. А.Ф. Можайского). Область научных интересов: обработка сигналов со сложной частотно-временной структурой в условиях априорной неопределенности относительно параметров сигналообразования. Число научных публикаций – 100. eremeeviu@yandex.ru; ул. Ждановская, 13, 197198, Санкт-Петербург, Россия; р.т.: +7(812)347-97-70; факс: +7(812)347-95-57.

Татарка Максим Васильевич – канд. техн. наук, докторант, Военно-космическая академия имени А.Ф. Можайского (ВКА им. А.Ф. Можайского). Область научных интересов: теория вероятностей и математическая статистика, распознавание образов, анализ больших объемов данных. Число научных публикаций – 22. maksimtbv@gmail.com; ул. Ждановская, 13, 197198, Санкт-Петербург, Россия; р.т.: +7(812)347-97-70; факс: +7(812)237-12-49.

Шуваев Федор Леонидович – канд. техн. наук, научный сотрудник, Военно-космическая академия имени А.Ф. Можайского (ВКА им. А.Ф. Можайского). Область научных интересов: теория вероятностей и математическая статистика, распознавание образов. Число научных публикаций – 8. cadetfed@mail.ru; ул. Ждановская, 13, 197198, Санкт-Петербург, Россия; р.т.: +7(812)347-97-70; факс: +7(812)237-12-49.

Цыганов Андрей Сергеевич – канд. техн. наук, старший преподаватель, Военно-космическая академия имени А.Ф. Можайского (ВКА им. А.Ф. Можайского). Область научных интересов: теория вероятностей и математическая статистика, распознавание образов, цифровая обработка сигналов. Число научных публикаций – 18. porudchik@mail.ru; ул. Ждановская, 13, 197198, Санкт-Петербург, Россия; р.т.: +7(812)347-97-70; факс: +7(812)237-12-49.

DOI 10.15622/ia.2020.19.6.7

I. EREMEEV, M. TATARKA, F. SHUVAEV, A. CYGANOV

COMPARATIVE ANALYSIS OF CENTRALITY MEASURES OF NETWORK NODES BASED ON PRINCIPAL COMPONENT ANALYSIS

Eremeev I., Tatarka M., Shuvaev F., Cyganov A. **Comparative Analysis of Centrality Measures of Network Nodes based on Principal Component Analysis**.

Abstract. The analysis of networks of a diverse nature, which are citation networks, social networks or information and communication networks, includes the study of topological properties that allow one to assess the relationships between network nodes and evaluate various characteristics, such as the density and diameter of the network, related subgroups of nodes, etc. For this, the network is represented as a graph – a set of vertices and edges between them. One of the most important tasks of network analysis is to estimate the significance of a node (or in terms of graph theory – a vertex). For this, various measures of centrality have been developed, which make it possible to assess the degree of significance of the nodes of the network graph in the structure of the network under consideration.

The existing variety of measures of centrality gives rise to the problem of choosing the one that most fully describes the significance and centrality of the node.

The relevance of the work is due to the need to analyze the centrality measures to determine the significance of vertices, which is one of the main tasks of studying networks (graphs) in practical applications.

The study made it possible, using the principal component method, to identify collinear measures of centrality, which can be further excluded both to reduce the computational complexity of calculations, which is especially important for networks that include a large number of nodes, and to increase the reliability of the interpretation of the results obtained when evaluating the significance node within the analyzed network in solving practical problems.

In the course of the study, the patterns of representation of various measures of centrality in the space of principal components were revealed, which allow them to be classified in terms of the proximity of the images of network nodes formed in the space determined by the measures of centrality used.

Keywords: Principal Component Analysis, Measure of Centrality, Graph, Clustering, Measure of Similarity

Eremeev Igor – Ph.D., Dr.Sci., Professor, Professor, Mozhaisky Military Space Academy. Research interests: difficult time-and-frequency structure signal processing under the conditions generation signal parameters prior indetermination. The number of publications – 100. eremeeviu@yandex.ru; 13, Zdanovskaya str., 197198, St. Petersburg, Russia; office phone: +7(812)347-97-70; fax: +7(812)347-95-57.

Tatarka Maxim – Ph.D., Doctoral Student, Mozhaisky Military Space Academy. Research interests: probability theory and mathematical statistics, pattern recognition, analysis of data. The number of publications – 22. maksimtbv@gmail.com; 13, Zdanovskaya str., 197198, St. Petersburg, Russia; office phone: +7(812)347-97-70; fax: +7(812)237-12-49.

Shuvaev Fedor – Ph.D., Researcher, Mozhaisky Military Space Academy. Research interests: probability theory and mathematical statistics, pattern recognition, analysis of data. The number of publications – 8. cadetfed@mail.ru; 13, Zdanovskaya str., 197198, St. Petersburg, Russia; office phone: +7(812)347-97-70; fax: +7(812)237-12-49.

Tsyganov Andrey — Ph.D., Senior Lecturer, Mozhaisky Military Space Academy. Research interests: probability theory and mathe-matical statistics, pattern recognition. The number of publications — 18. porudchik@mail.ru; 13, Zdanovskaya str., 197198, St. Petersburg, Russia; office phone: +7(812)347-97-70; fax: +7(812)237-12-49.

References

1. Bonchi F., De Francisci G., Riondato M. Centrality Measures on Big Graphs: Exact, Approximated, and Distributed Algorithms. Proceedings of the 25th International Conference Companion on World Wide Web. 2016. pp. 1017–1020.
2. Shherbakova N.G. [Centrality measures in networks]. *Problemy informatiki – Informatics problems*. 2015. vol. 1. pp. 18–30. (In Russ.).
3. Beredihin S.V., Ljapunov V.M., Shherbakova N.G. [Measure of importance of scientific periodicals – «Centrality by mediation»]. *Problemy informatiki – Informatics problems*. 2014. vol. 3. pp. 53–63. (In Russ.).
4. Judina M.N. [Nodes in social networks: measures of centrality and role in network processes]. *Omskij nauchnyj vestnik – Omsk Scientific Bulletin*. 2016. № 4. pp. 161–165. (In Russ.).
5. Brandes U., Borgatti S., Freeman L. Maintaining the duality of closeness and betweenness centrality. *Social Networks*. 2016. vol. 44. pp. 153–159.
6. Minoo A. et al. A Systematic Survey of Centrality Measures for Protein-Protein Interaction Networks. *BMC Systems Biology*. 2018. vol. 12. no. 1. pp. 80.
7. Chen P-Y., Choudhury S., Hero A., Multi-centrality graph spectral decompositions and their application to cyber intrusion detection. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. pp. 4553–4557.
8. Lu B., Sun H., Harris P., Xu M. Shp2graph: Tools to Convert a Spatial Network into an Igraph Graph in R. *ISPRS Int. J. Geo-Inf.* 2018. vol. 7. pp. 293.
9. Csardi G, Nepusz T. The IGRAPH software package for complex network research // InterJournal, Complex Systems. 1695. 2006. vol. 1695. no. 5. pp. 1–9.
10. Shuvaev F.L., Tatarka M.V. [Analysis of the dynamics of measures of centrality of mathematical models of random graphs]. *Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki – Scientific and technical bulletin of information technologies, mechanics and optics*. 2020. Issue 20. vol. 2. pp. 249–256. (In Russ.).
11. Shuvaev F.L., Tatarka M.V. [Analysis of mathematical models of random graphs used in the simulation of information and communication networks]. *Vestnik Sankt-Peterburgskogo universiteta GPS MChS Rossii – Bulletin of St. Petersburg University State Fire Service of the Ministry of Emergencies of Russia*. 2020. vol. 2. pp. 67–77. (In Russ.).
12. Van Mieghem P. et al. Spectral graph analysis of modularity and assortativity. *Phys. Rev.* 2010. vol. 82. no. 5. P. 056113.
13. Barzel B., Biham O. Quantifying the connectivity of a network: the network correlation function method. *Phys. Rev*. 2009. vol. 80. pp. 046104.
14. Barabasi A. Network Science. Cambridge university press. 2016. 453 p.
15. Watts D., Strogatz H. Collective dynamics of «Small-world» networks. *Nature.* 1998. vol. 393. pp. 440–442.
16. Hartmann A., Mézard M. Distribution of diameters for Erdős-Rényi random graphs. *Phys. Rev*. 2018. vol. 97. no. 3. pp. 032128.
17. Le C., Levina E., Vershynin R. Concentration and regularization of random graphs. *Random Structures&Algorithms*. 2017. vol. 51. no. 3. pp. 538–561.
18. Gibson H., Vickers P. Using adjacency matrices to lay out larger small-world networks. *Applied soft computing*. 2016. vol. 42. pp. 80–92.
19. Jalili M. et al. CentiServer: A Comprehensive Resource, Web-Based Application and R Package for Centrality Analysis. *PLoS ONE*. 2015. vol. 10. no. 11. pp. 0143111.
20. Oldham, S. et al. Consistency and differences between centrality measures across distinct classes of networks. *PLoS ONE*. 2019. vol. 14. no. 7. pp. 0220061.
21. Bloch F., Jackson M., Tebaldi P. Centrality measures in networks. SSRN. 2016. 42 p.
22. Lê S., Josse J., Husson F. FactoMineR: A Package for Multivariate Analysis. *Journal of Statistical Software.* 2008. vol. 25. no. 1. pp. 1–18.
23. Depaolini M., Ciucci D., Calegari S., Dominoni M. External Indices for Rough Clustering. *Lecture Notes in Computer Science*. 2018. vol. 11103. pp. 378–391.
24. White S., Smyth P. Algorithms for estimating relative importance in networks. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003. pp. 266–275.