УДК 519.1, 316.33

ПОСТРОЕНИЕ И ИССЛЕДОВАНИЕ СТРУКТУРЫ СЛОЖНОЙ СЕТИ YOUTUBE - КАНАЛОВ

Дидоренко Алексей Викторович[1], Прогулова Татьяна Борисовна[2]

[1]Магистрант;
Государственный университет «Дубна»;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: alx.did.00@gmail.com.

[2]Кандидат технических наук, доцент;
Государственный университет «Дубна»;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: progulova@yahoo.com.

Цель работы – построение и анализ структурных особенностей сложной сети YouTube – каналов в рамках теории сложных сетей. Для изучения топологии и динамики сети в ходе работы были собраны данные для построения пяти карт фрагмента сети YouTube-каналов с интервалом около 2 недель. Были вычислены и проанализированы базовые характеристики сети: диаметр, радиус, коэффициент кластеризации, корреляции степеней и т.д. Была изучена динамика роста степеней узлов, показано, что сеть является растущей с нелинейным предпочтительным присоединением с показателем, равным 0.7. Особое внимание уделено изучению топологических особенностей сети, включая структуру сообществ. Проанализирована тематическая направленность каналов, образующих сообщества.Полученные результаты позволяют генерировать сети, статистически подобные сети каналов YouTube, и исследовать в них особенности распространения информации.

Ключевые слова: теория сложных сетей, анализ социальных сетей, *YouTube*.

Для цитирования:
Дидоренко А. В., Прогулова Т. Б. Построение и исследование структуры сложной сети YouTube - каналов// Системный анализ в науке и образовании: сетевое научное издание. 2022. № 1. С. 77–90. URL : http://sanse.ru/download/462.

CONSTRUCTING AND ANALYSIS THE STRUCTURE OF A COMPLEX YOUTUBE-CHANNELS NETWORK

Didorenko Aleksei V.[1], Progulova Tatyana B.[2]

[1]Master's Degree student;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: alx.did.00@gmail.com.

[2]PhD in Physical and Mathematical Sciences, associate professor;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: progulova@yahoo.com.

The purpose of the work is to construct and study the structure features of the YouTube-channels network within the theory of complex networks. To study the topology and dynamics of the network in the course of the work the data were collected to construct seven maps of a fragment of the YouTube-channels network with different intervals. The basic characteristics of the network were calculated and analyzed: diameter, radius, clustering coefficient, degree correlations, etc. The dynamics of node degree growth was studied, and it was shown that the network is growing with nonlinear preferential attachment with an index equal to 0.7. Special attention was paid to the study of topological features of the network, including the structure of communities. The thematic orientation of the channels forming the communities was analyzed.

The results allow us to generate networks statistically similar to the YouTube channel network and study the peculiarities of information spreading in them.

Keywords: complex networks theory, social network analysis, YouTube.

For citation:
Didorenko A. V., Progulova T. B. Constructing and analysis the structure of a complex YouTube-channels network. System Analysis in Science and Education, 2021;(1): 77–90(In Russ). Available from: http://sanse.ru/download/462.

Введение

В последние годы сложно-сетевой подход, использующий понятия и методы математики, физики, информатики и социологии, сформировался в универсальный инструмент исследования реальных сложных систем, моделируемых как сложные сети [1].

Возникшая в начале XXI века наука о сложных сетях позволяет описывать существующие в природе и обществе сети, обладающие нетривиальными топологическими свойствами, которые не являются типичными для случайных графов или регулярных решеток. Как правило, такая сеть содержит малое число узлов большой степени – хабов (от английского *hub* – ядро, концентратор), существенно определяющих её свойства и возможности функционирования. При этом оказалось, что большинство реальных сетей (биологических, технических, социальных) являются сложными [2].

Наглядным примером сложной и социально значимой системы является видеохостинг *YouTube*. Сервис содержит около 2 миллиардов каналов и обеспечивает возможность подписываться на понравившиеся каналы, проводить онлайн-трансляции; загружать, оценивать и комментировать видеоролики, а также делиться ими.

На протяжении всего своего существования *YouTube* достаточно сильно эволюционировал. Покупка компанией *Google*, блокировка и локализация в различных странах мира, коммерциализация, внедрение системы защиты авторского права – все эти этапы развития видеохостинга отразились в эволюции структуры *YouTube* как сложной системы [3].

По мере возрастания популярности *YouTube*, возрастал и интерес к изучению его структуры. Так, Джон Паолильо с соавторами исследовали историю контента *YouTube* как меняющуюся со временем сеть отношений между каналами, выраженную в плейлистах «Нравится/избранное». В своей работе они обсуждают, как внешние и внутренние факторы могут изменять тип контента во всей сети и отдельно выделенных кластерах [4]. Малек Джебабли с соавторами проводили сравнительный анализ сети *YouTube* на уровне пользователей (каналов) и групп (сообществ пользователей) и показали, что два разных типа рассматриваемых сетей имеют схожую структуру [5]. Результаты этого исследования позволяют лучше понять отношения между мезоскопическими и макроскопическими свойствами онлайновых социальных сетей как с топологической, так и с функциональной точки зрения.

Тем не менее, несмотря на популярность видеохостинга *YouTube*, он до сих пор остаётся малоизученным. И это при том, что его социальная значимость, как совершенного ресурса обмена информацией разного рода, постоянно возрастает. Функционирование видеохостинга привлекает внимание обыкновенных пользователей и организаций. Интригующей остаётся роль *YouTube* как информационного оружия в избирательных компаниях, брекзите, формировании общественного мнения или фейков, и многом другом. Все вышеперечисленные процессы на уровне каналов протекают с учетом структуры сети подписок, но под управлением системы рекомендаций, принцип работы которой досконально не известен. Чтобы изучать, понимать и управлять процессами в сети *YouTube*, особенно с учетом этой неопределенности, необходимо как можно более адекватно представлять структуру базовой сети и закономерности ее эволюции. Попытка такого исследования предпринята рамках данной работы.

Метод исследования

YouTube сформировал гигантскую сеть, – изучение ее возможно только на уровне подсетей (выборок). Полная карта *YouTube* доступна только *Google* / *YouTube*, а данные для отдельных видео доступны

только через функции поиска и просмотра и не могут быть отобраны действительно случайным образом (например, вследствие спонсируемых функций поиска и алгоритмов рекомендаций по видео) [6]. Поэтому в данной работе анализировалась структура сети *YouTube* на уровне каналов: узлам соответствовали *YouTube*-каналы, а направленные связи – отношениям подписки.

Для проведения исследования необходимо было построить выборки с использованием просмотра и сканирования *API* общедоступных данных *Google* / *YouTube* с последующем извлечением сетевых данных. Для каждой из сетевых карт вычислялись основные сетевые характеристики и определялась структура сообществ, из анализа временной последовательности сетевых карт извлекалась информация о закономерностях эволюции сети каналов *YouTube*.

Построение выборок

Данные о сети каналов *YouTube* (ее подсети) можно получить из репозитория, специализирующегося на анализе сложных сетей, например, *Network Data Repository* [7]. Однако, такие данные, во-первых, достаточно устаревшие, и во-вторых, позволяют исследовать только статический граф (вершины и направленные ребра), не давая возможности получить информацию о динамике изменений сети (правилах эволюции), а также природе сообществ. Поэтому для построения выборок был использован сканер сетевых данных, позволяющий собрать актуальную и более полную информацию о *YouTube*-каналах и их подписчиках.

Эмпирические данные собирались с помощью веб-приложения *YouTube Data Tools*. Этот инструмент позволяет сканировать данные платформы *YouTube* через *YouTube API v*3 и формировать протокол в стандартных форматах файлов для дальнейшего анализа в других программных пакетах [8].

При сборе данных сканирование начиналось с 50 самых популярных *YouTube*-каналов, то есть таких, которые имеют наибольшее число подписчиков. Для изучения закономерностей эволюции сети сканирование при глубине сканирования, равной «2», проводилось 5 раз с временным промежутком в 2 недели с 15 марта по 15 мая 2021 года. Дополнительно для изучения вопроса о зависимости структуры подсети от глубины сканирования аналогичная сеть была собрана с помощью глубины «3» в августе 2021 года. Наконец, для исследования зависимости структуры подсети от локализации в мае 2021 года были аналогичным образом собраны данные по русскоязычным каналам.

Глубина сканирования показывает насколько «глубоко» должен работать сканнер (см. рис. 1).

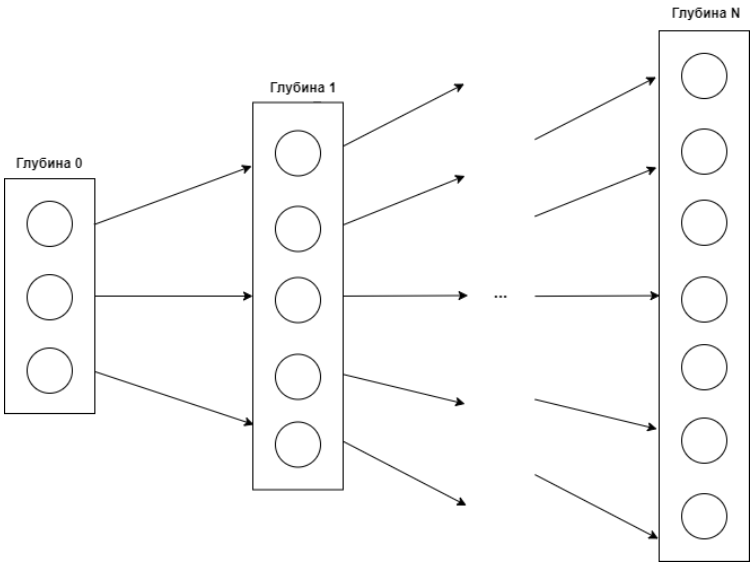

Рис. 1. Схема работы сборщика данных

Стартуя от начального узла при каждом значении глубины сканер добавляет полученную информацию о каналах, на которые текущий узел подписан, в текстовый *gdf*-файл. Данный файл содержит описание сети, каждая строка списка представляет ребро, задаваемое уникальным идентификатором

входного и выходного канала, а также логическим значением, указывающим, является ребро направленным или нет (см. рис. 2) [8].

```
16104 edgedef>node1 VARCHAR,node2 VARCHAR,directed BOOLEAN
16105 UCpEhnqL0y41EpW2TvWAHD7Q,UC6uPyCHIzAx4pb4Pd0dj_mw,true
16106 UCpEhnqL0y41EpW2TvWAHD7Q,UCAi46BZsJjXCmOYcfbM-kqQ,true
16107 UCpEhnqL0y41EpW2TvWAHD7Q,UCHu48NlukyWGqjh3DUKcBmA,true
16108 UCpEhnqL0y41EpW2TvWAHD7Q,UCcIPw0eP8KQPUXh5rK4wp3A,true
16109 UCpEhnqL0y41EpW2TvWAHD7Q,UCOQNJjhXwvAScuELTT_i7cQ,true
16110 UCpEhnqL0y41EpW2TvWAHD7Q,UCnpaBg-u_kHwzuPyaMcyJ0w,true
16111 UCpEhnqL0y41EpW2TvWAHD7Q,UCQmxcMxjYcBM5Pel4qUW2hA,true
16112 UCpEhnqL0y41EpW2TvWAHD7Q,UCkRfArvrzheW2E7b6SVT7vQ,true
16113 UCpEhnqL0y41EpW2TvWAHD7Q,UCw7xjxzbMwgBSmbeYwqYRMg,true
16114 UCpEhnqL0y41EpW2TvWAHD7Q,UCF_cPp4EXD_BFUkL8HHJ8BA,true
16115 UCpEhnqL0y41EpW2TvWAHD7Q,UCtahuT3COFyZ8xs4TDAfIjw,true
16116 UCpEhnqL0y41EpW2TvWAHD7Q,UCBIT1FSJW6yTlzqK-31FDWg,true
16117 UCpEhnqL0y41EpW2TvWAHD7Q,UCBz9qbqujcjIavFmBJW2oZw,true
16118 UCpEhnqL0y41EpW2TvWAHD7Q,UC25he-g2TMwFx2AnEiCkrzA,true
```

Рис. 2. Структура данных, собираемая сканером, соответствующая списку рёбер

Построение сети и последующий ее анализ проводились с помощью библиотек *tuber* (удобный клиент для работы с *YouTube API*) и *igraph* (инструментарий для анализа сложных сетей) языка статистического программирования *R* в среде разработки *RStudio*.

Анализ структуры сложной сети YouTube – каналов

Общие характеристики

После построения сетей *YouTube*-каналов вычислялись общие характеристики (количество вершин и ребер), средние степени, значение коэффициента кластеризации, метрические характеристики, распределение степеней. Они представлены в таблице 1. Все сети являются связными.

Таблица 1. Основные характеристики построенных сетей. Сети 1-5 собраны при глубине сканирования «2», сеть 6 – при глубине «3». Сеть 7 получена сканированием русскоязычных каналов

	Сеть 1	**Сеть 2**	**Сеть 3**	**Сеть 4**	**Сеть 5**	**Сеть 6**	**Сеть 7**
Количество узлов	15949	15997	15976	16037	16095	402093	10383
Количество ребер	305703	306533	303974	305468	306477	8670620	142198
Направленная сеть	Да	Да	Да	Да	Да	Да	Да
Диаметр	14	14	14	14	14	15	15
Радиус	2	2	2	2	2	6	2
Среднее расстояние	4.26	4.26	4.26	4.26	4.26	5.08	4.84
Коэффициент кластеризации	0.067	0.067	0.066	0.065	0.066	0.014	0.102
Коэффициент кластеризации случайной сети того же размера	0.0024	0.0024	0.0024	0.0024	0.0024	0.0001	0.003
Средняя степень	38.3	38.3	38.1	38.1	38.1	43.1	27.4
Средняя полустепень исхода/захода	19.2	19.2	19	19.1	19	21.6	13.7

Для сетей 1-5, полученных при глубине сканирования «2», и по сути описывающих эволюцию практически одной и той же подсети каналов *YouTube*, основные характеристики, как и можно было

ожидать, очень близки по значениям. Можно видеть, что сеть растет даже за такой небольшой промежуток времени. Неожиданно, что третья карта является наименьшей по количеству вершин и рёбер среди других карт. Возможно, причиной такого явления является то, что на момент сбора карты третьей сети происходили различные блокировки доступа к платформе *YouTube*. Ожидаемо, за всё время исследования в сети не изменялись диаметр и радиус. Это, по всей видимости, следствие заданной глубины сканирования «2», хотя инициация просмотра от 50 наиболее популярных каналов могла привести и к большим значениям радиуса сети. Незначительно менялась средняя длина пути.

Значения коэффициента кластеризации, отражающего вероятность того, что соседи узла связаны между собой (плотность треугольников), на первый взгляд, не велики – $0.065 - 0.067$, но сравнение со значениями коэффициентов кластеризации для случайных сетей Эрдеша-Реньи с тем же количеством вершин и ребер (или при рандомизации исходного графа), дает 0.0024 (почти в 30 раз меньше). Таким образом, исследуемая сеть является высоко кластеризованной, что вместе с небольшими значениями среднего расстояния между узлами указывает на ее мало-мировой характер [9].

Сеть 6, полученная при глубине сканирования «3», имеет более высокие значения базовых характеристик, чем у предыдущих сетей. Значение коэффициента кластеризации – 0.014 – существенно отличается от значения коэффициента кластеризации для случайной сети Эрдеша-Реньи с тем же количеством вершин и ребер, равного 0.0001 (в 140 раз меньше). При увеличении числа вершин по сравнению с сетями 1-5 в ~25 раз среднее расстояние между вершинами этой карты сети увеличилось менее чем на 20%, что опять-таки подтверждает мало-мировое свойство.

Базовые характеристики вычислялись также для сети 7, полученной из сканирования русскоязычных каналов. Несмотря на то, что это совсем другая подсеть сети каналов *YouTube*, свойства ее очень близки к сетям, рассмотренным выше.

Таким образом, можно утверждать, что сеть каналов *YouTube* в целом, является значительно кластеризованной и мало-мировой.

Распределение степеней

Одной из важнейших характеристик структуры сложной сети является распределение степеней. Поскольку изучаемые сети ориентированные, были построены распределения полных степеней, полустепеней захода и полустепеней исхода. На рисунках 3-5 представлены распределения для сети, полученной с при глубине сканирования «3».

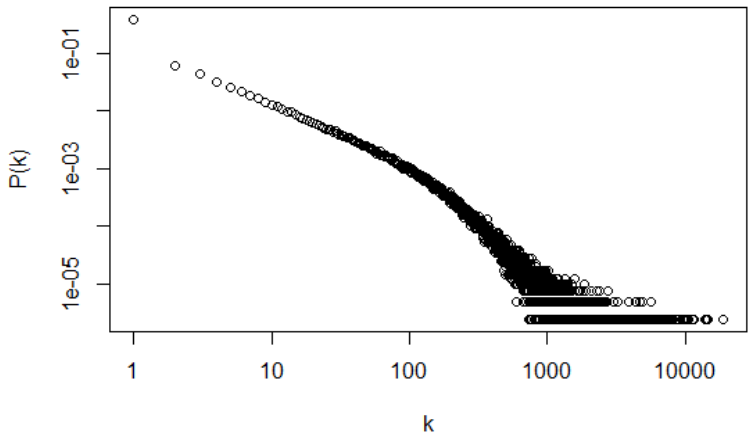

Рис. 3. Распределение степеней сети 6 в log-log масштабе

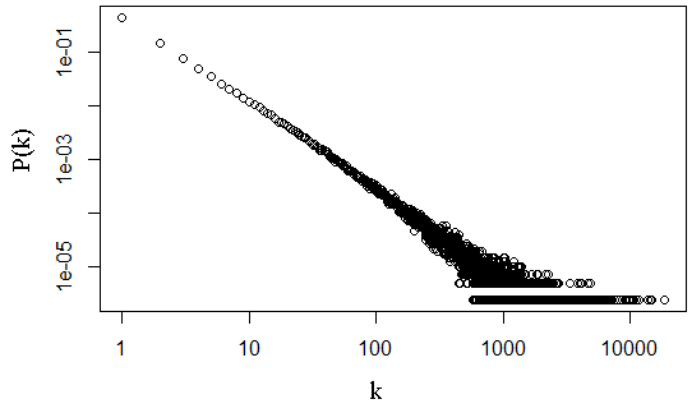

Рис. 4. Распределение полустепеней захода сети 6 в log-log масштабе

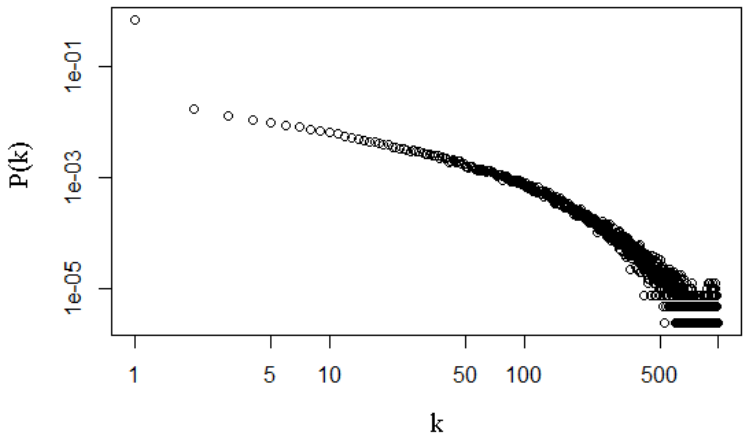

Рис. 5. Распределение полустепеней исхода сети 6 в log-log масштабе

Форма распределений в *log-log* масштабе с прямолинейным участком указывает на то, что исследуемая сеть является безмасштабной, с распределением степеней, описываемым степенной функций $P(k) \sim k^{-\gamma}$ [9]. При малых степенях есть отклонения от прямой, это может быть связано с малой глубиной сканирования во время сбора данных. Значение показателя степени γ были получены с помощью метода максимального правдоподобия (1):

$$\gamma = 1 + n\left[\sum_{i=1}^{n} ln\frac{k_i}{k_{min}}\right]^{-1}, \qquad (1)$$

где γ – оценка показателя степени, n – число узлов в сети, k_{min} – минимальная степень, k_i – степень i-го узла [10]. Значения x_{min} при аппроксимации выбирались так, чтобы описать прямолинейную часть распределения и толстый хвост. Результаты приведены в таблице 2.

Таблица 2. Оценки показателя степени для распределений степеней сетей 1 – 7

Режим	**Значения**	**Сеть 1**	**Сеть 2**	**Сеть 3**	**Сеть 4**	**Сеть 5**	**Сеть 6**	**Сеть 7**
Полная степень	k_{min}	32	33	32	32	32	53	24
	γ	2.01 ± 0.02	2.01 ± 0.02	2.003 ± 0.02	2.002 ± 0.02	2.003 ± 0.02	2.007± 0.004	2.01 ± 0.02

Режим	Значения	Сеть 1	Сеть 2	Сеть 3	Сеть 4	Сеть 5	Сеть 6	Сеть 7
Полустепень захода	k_{min}	15	15	15	15	15	188	9
	γ	2.01 ± 0.02	2.01 ± 0.02	2.01 ± 0.02	2.01 ± 0.02	2.01 ± 0.02	2.001± 0.014	2.01 ± 0.02
Полустепень исхода	k_{min}	41	41	41	41	41	39	29
	γ	2.01 ± 0.02	2.01 ± 0.02	2.01 ± 0.02	2.01 ± 0.02	2.01 ± 0.02	2.005 ± 0.004	2.02 ± 0.03

Таким образом, распределения степеней всех анализируемых сетей описываются степенным законом $P(k)\sim k^{-\gamma}$ с показателем степени $\gamma \approx 2$. Такой характер распределения указывает на то, что сеть является растущей с предпочтительным присоединением (то есть новые поступающие в сеть вершины соединяются с уже присутствующей в сети вершиной с вероятностью $\Pi(k)$ тем большей, чем больше ее степень). В базовой модели безмасштабных сетей – модели Барабаши-Альберт – $\Pi(k)\sim k$, и это приводит к сетям с распределением степеней с $\gamma = 3$. Полученное для сетей *YouTube* значение $\gamma \approx 2$ указывает на присутствие дополнительных механизмов эволюции сети: удаление узлов, создание новых связей между «старыми» узлами, старение узлов и, наконец, нелинейное предпочтительное присоединение [9].

Изучение характера предпочтительного присоединения

Для исследования характера предпочтительного присоединения использовалась следующая методика. Рассмотрим сеть, для которой у нас есть две разные карты: первая, взятая в момент времени t, а вторая в момент времени $t + \Delta t$ (см. рис. 6*a*). Для узлов, которые изменили свою степень в течение Δt, измеряем изменение степени Δk_i. Согласно предпочтительному присоединению, относительное изменение $\Delta k_i/\Delta t$ должно быть пропорционально вероятности $\Delta k_i/\Delta t \sim \Pi(k_i)$, обеспечивая функциональную форму предпочтительного присоединения. На практике полученная кривая $\Delta k_i/\Delta t$ может быть зашумленной. Чтобы уменьшить этот шум, измеряют кумулятивную функцию предпочтительного присоединения (2):

$$\pi(k) = \sum_{k_i=0}^{k} \Pi(k_i). \quad (2)$$

При отсутствии предпочтительного присоединения имеем $\Pi(k_i) = const$, следовательно, кумулятивная функция $\pi(k)\sim k$. Если имеет место линейное предпочтительное присоединение, то есть если $\Pi(k_i)\sim k$, ожидается, что кумулятивная функция $\pi(k)\sim k^2$ (см. рис. 6*b*) [9].

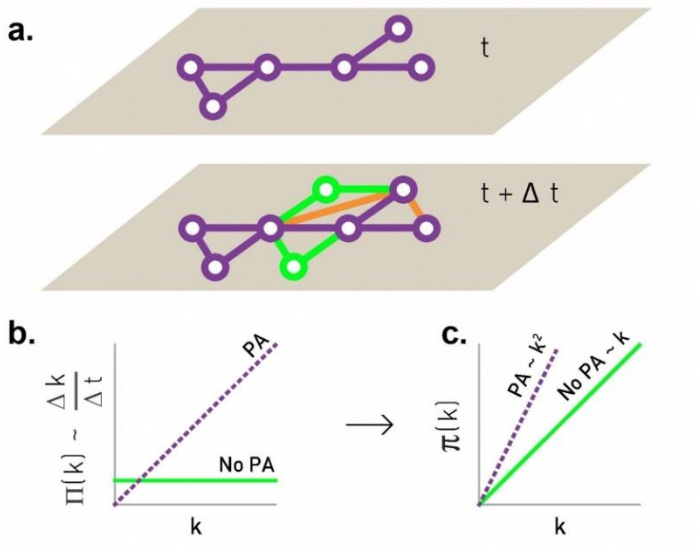

Рис. 6. Измерение характера предпочтительного присоединения

Для оценки показателя предпочтительного присоединения использовались пять карт, на основе которых вычислялась кумулятивная зависимость частоты изменения степени вершин, которая представлена на рисунке 7. Синие точки соответствует полученной зависимости, а оранжевая прямая линия соответствует зависимости в случае линейного предпочтительного присоединения. Сравнение двух зависимостей указывает на присутствие нелинейного предпочтительного присоединения в исследуемой сети с показателем нелинейности меньше 1, а именно 0.7. Стоит отметить, что нелинейность также указывает на дополнительные механизмы, влияющие на эволюцию сети [9].

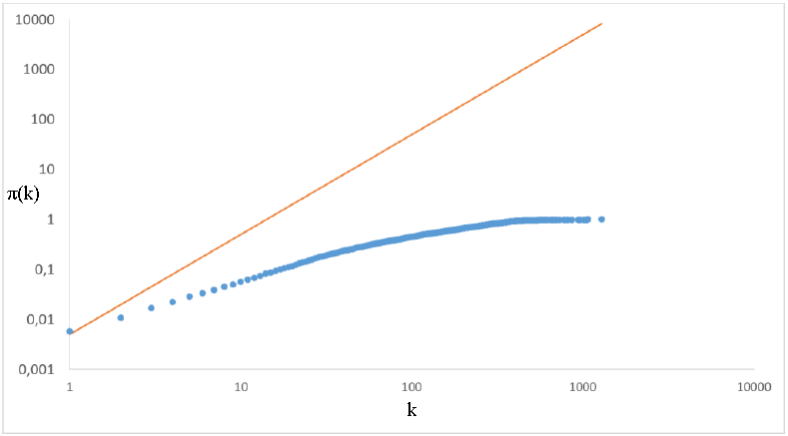

Рис. 7. Зависимость $\pi(k)$, вычисленная для фрагмента сети каналов YouTube. Синие точки соответствуют эмпирической зависимости, прямая линия – оценочная – соответствует значению показателя предпочтительного присоединения $\alpha = 1$

Были сгенерированы растущие сети с нелинейным предпочтительным присоединением $\Pi(k_i) \sim k^{0.7}$. На рисунке 8 показано распределение степеней, значение показателя γ составило $\gamma = 2.010 \pm 0.007$. Сходство топологий эмпирической и сгенерированных сетей позволяет рассматривать нелинейное предпочтительное присоединение как ключевой механизм эволюции сети, позволяющий генерировать сети, статистически схожие с исследуемой сетью и изучать на них процессы распространения информации.

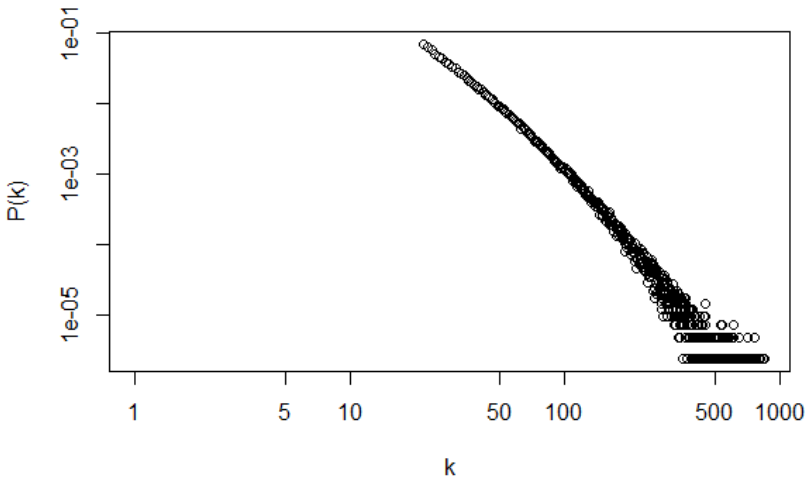

Рис. 8. Распределение степеней сгенерированной сети, $\gamma = 2.010 \pm 0.007$

Корреляции степеней

Для определения наличия и характера корреляций степеней вершин вычислялся коэффициент ассортативности, отражающий тенденцию узлов сети соединяться с другими узлами, близкими по степени. Для сетей, собранных при глубине сканирования «2», коэффициент ассортативности равен -0.058, для сети, полученной при глубине «3», равен -0.096. Для более детального анализа корреляций также строилась зависимость средней степени ближайших соседей вершины от её степени $\langle k_{nn} \rangle (k)$. На рисунках 9-10 приведены графики $\langle k_{nn} \rangle (k)$ для сетей, полученные при глубине «2».

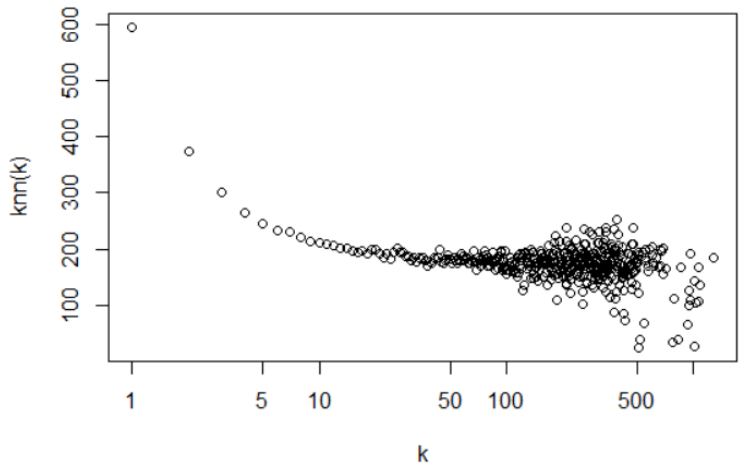

Рис. 9. Зависимость средней степени ближайших соседей вершины от её степени $\langle k_{nn} \rangle (k)$ для сети 5

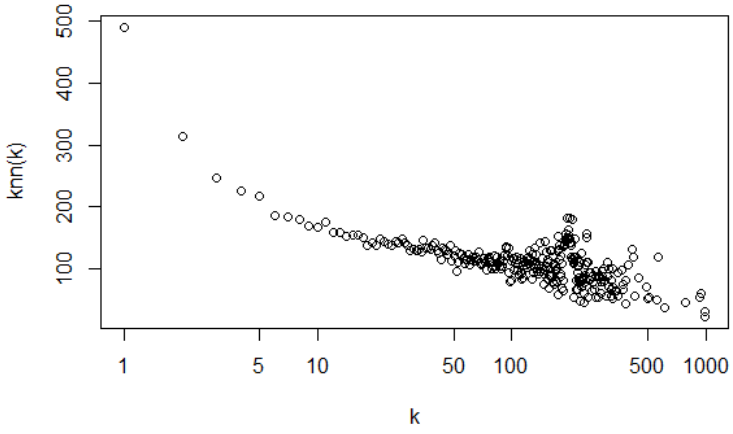

Рис. 10. Зависимость средней степени ближайших соседей вершины от её степени $\langle k_{nn} \rangle (k)$ для сети 7 (русскоязычная подсеть)

Отрицательное значение коэффициента ассортативности и убывающий характер зависимостей указывают на то, что сети слабо-дисассортативны, то есть вершины с большой степенью имеют тенденцию соединяться с вершинами малой степени и наоборот. Однако известно, что сети с ярко выраженной структурой сообществ, что можно ожидать для сети каналов *YouTube*, как правило, ассортативны. Наблюдаемая дисассортативность может быть вызвана конфликтом между корреляциями степеней и масштабно-инвариантным характером распределения степеней (как следствие того, что сеть рассматривается как простая, т.е. без кратных связей между узлами) [9].

Для подробного изучения характера корреляций степеней вершин детальнее анализировалась сеть, полученная при глубине сканирования «3», график зависимости $\langle k_{nn} \rangle (k)$ которой имеет аномаль-

ную форму. Есть растущая часть графика при небольших значения k, указывающая на ассортативность, и убывающая часть графика при больших k, свидетельствующая о дисассортативном характере (см. рис. 11).

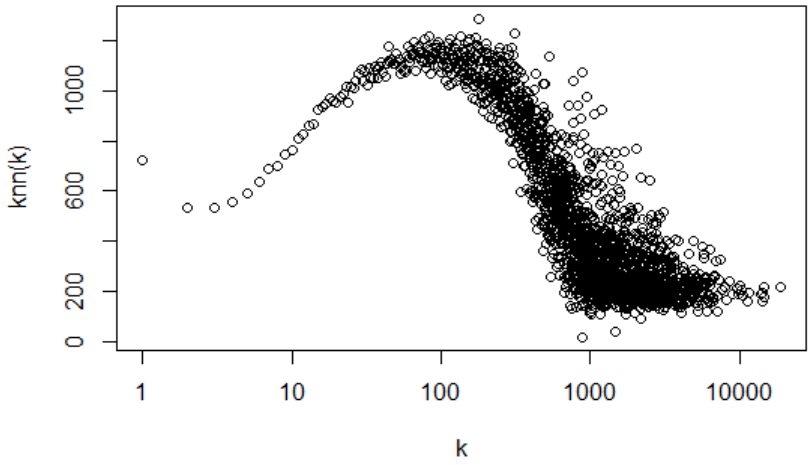

Рис. 11. Зависимость средней степени ближайших соседей вершины от её степени $\langle k_{nn}\rangle(k)$ для сети 6

В сети с корреляциями степеней $e_{kk'}$ ожидаемое количество связей между узлами со степенями k и k' равно $E_{kk'} = e_{kk'}\langle k\rangle N$, где $\langle k\rangle$ – средняя степень сети, а N – число узлов в сети. Возьмем два крупнейших узла сети, они имеют степени k = 18917 и k' = 14762. В предположении нейтральной сети и с учетом значений $\langle k\rangle$ и N для сети 6 (см. рис. 3) имеем (3):

$$E_{kk'} = \frac{kp_k k'p_{k'}}{\langle k\rangle}N = \frac{18917*14762}{43.1*402093} \approx 16. \quad (3)$$

Таким образом, учитывая размер этих двух хабов, они должны быть соединены друг с другом примерно 16 связями, чтобы соответствовать нейтральной природе сети. Тем не менее, в простой сети мы можем иметь только одну связь между ними, что и вызывает конфликт между корреляциями степеней и масштабно-инвариантным свойством. Тот факт, что не разрешены кратные связи (представление простой сети), делает сеть дисассортативной – явление, называемое структурной дисассортативностью. Для малых k и k' ожидаемое количество связей $E_{kk'}$ также мало, то есть ожидается, что между двумя узлами будет меньше одной связи. Мультисвязи ожидаются только для узлов, степень которых превышает некоторый порог k_s. Можно оценить k_s, называемое структурным отсечением, для изучаемой сети (4):

$$k_s = (\langle k\rangle N)^{\frac{1}{2}} \approx 4124, \quad (4)$$

Чтобы понять последствия структурного отсечения, нужно определить, есть ли в сети узлы, степень которых превышает k_s. Для этого мы сравниваем структурное отсечение k_s с естественным ограничением на наибольшую степень в сети заданного размера со степенным распределением степеней (5):

$$k_{max} \sim N^{\frac{1}{\gamma-1}} \approx 397240. \quad (5)$$

Из сопоставления k_s и k_{max}, можно сделать вывод о присутствии в сети структурной дисассортативности. Иными словами, узлы, чья степень находится между k_s и k_{max}, могут нарушать $E_{kk'} > 1$, то есть сеть имеет меньше связей между узлами, чем предсказано. Поэтому такие сети проявляются как дисассортативные.

Для определения, являются ли корреляции, наблюдаемые в сети, следствием структурной дисассортативности или они генерируются каким-то неизвестным процессом, который приводит к корреляциям степеней, использовалась сохраняющая степень рандомизация. Если в результате зависимости $\langle k_{nn} \rangle(k)$ реальной и рандомизированной сетей неразличимы, то все корреляции, наблюдаемые в реальной системе, являются структурными и полностью объясняются распределением степеней. Иначе, если зависимость $\langle k_{nn} \rangle(k)$ рандомизированной сети не показывает корреляции степеней, в то время как для реальной сети показывает, то существует некоторый неизвестный процесс, который генерирует наблюдаемые корреляции степени [9]. Сравнение $\langle k_{nn} \rangle(k)$ для сети 6 и соответствующей рандомизированной сети показано на рисунке 12.

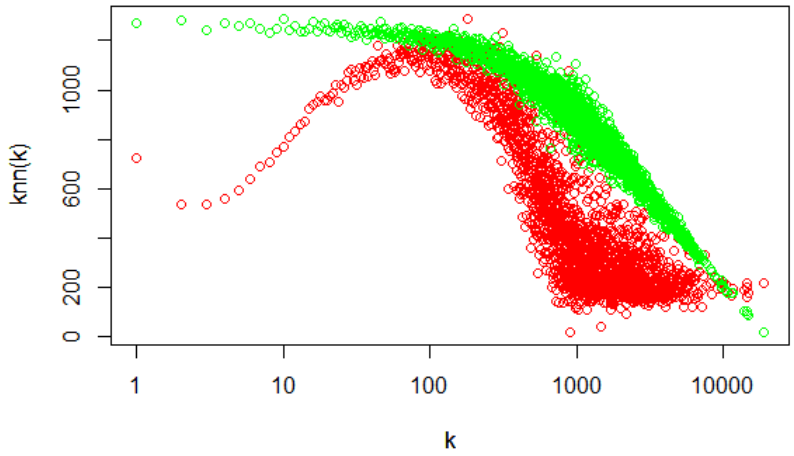

Рис. 12. Зависимость средней степени ближайших соседей вершины от её степени $\langle k_{nn} \rangle(k)$. Красная зависимость соответствует эмпирической сети 6, а зелёная – рандомизированной сети

Ассортативность в левой части, при небольших значениях k разрушилась, т.е. она присуща сети, что и является типичным для сетей со структурой сообществ, а дисассортативность при больших k, очевидно, является структурной.

Определение структуры сообществ и их интерпретация

Еще одной структурной особенностью, существенно влияющей на процессы в сети, являются сообщества. Для выделения сообществ использовался набор алгоритмов, реализованных в библиотеке *igraph*.

При изучении сообществ в сетях необходима метрика, чтобы объективно оценить, насколько «хорошим» является конкретное разделение сети на сообщества. Одной из таких метрик является мера модульности, указывающая насколько отличается локальная плотность связей в подсетях, определенных этим разделением, от ожидаемой плотности в случайной сети. Более высокое значение модульности подразумевает лучшее разделение. Значение модульности, равное 0, соответствует случаю, когда вся сеть рассматривается как одно сообщество [9].

Для изучаемых сетей наилучшие значения модульности даёт алгоритм *Louvain*. Этот алгоритм хорошо работает для больших сетей, особенно с высоким коэффициентом кластеризации и распределением степеней с толстыми хвостами, как это имеет место в сети каналов *YouTube*. Он выполняет так называемую агломеративную иерархическую кластеризацию, в которой узел назначается кластеру, если это максимизирует модульность сети, и продолжает до тех пор, пока либо не останется ни одного не назначенного сообществам узла, либо модульность не может быть увеличена дальше. Кластеризация по модульности несовершенна: иногда она выявляет несуществующие связи между кластерами на основе слабых ложноположительных связей и имеет тенденцию давать большое количество кластеров в разреженных сетях [11]. Тем не менее, она хорошо работает для обнаружения четко определенных кластеров в больших сетях, как ожидалось в случае *YouTube*.

Для каждой сети фиксировалось количество полученных сообществ и значение модульности (см. табл. 3).

Таблица 3. Результаты анализа структуры сообществ в сети

	Глубина 2	**Глубина 3**	**Русскоязычный сектор**
Количество сообществ	12	62	13
Модульность	0.4	0.5	0.5

Для понимания природы сообществ были определены соответствующие им тематики. Для этого извлекались и анализировались жанры видеороликов для каждого из каналов из больших сообществ, найденных алгоритмом *Louvain* (см. рис. 13-15). Из диаграмм видно, что в исследуемых сетях главные жанры контента носят развлекательный характер, что подтверждает то, что *YouTube* по большей части был и остается развлекательной платформой. Также среди жанров русскоязычного сектора были обнаружены сообщества, которые сконцентрированы на кинематографе. Возможно, это издержки более слабого соблюдения авторских прав в этом секторе.

Рис. 13. Жанры видеороликов в сообществах сети, полученной с помощью глубины «2»

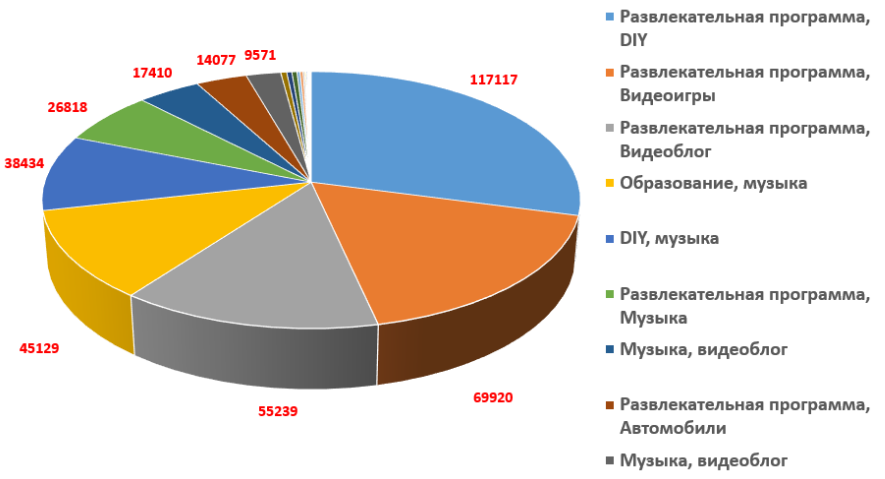

Рис. 14. Жанры видеороликов в сообществах сети, полученной с помощью глубины «3»

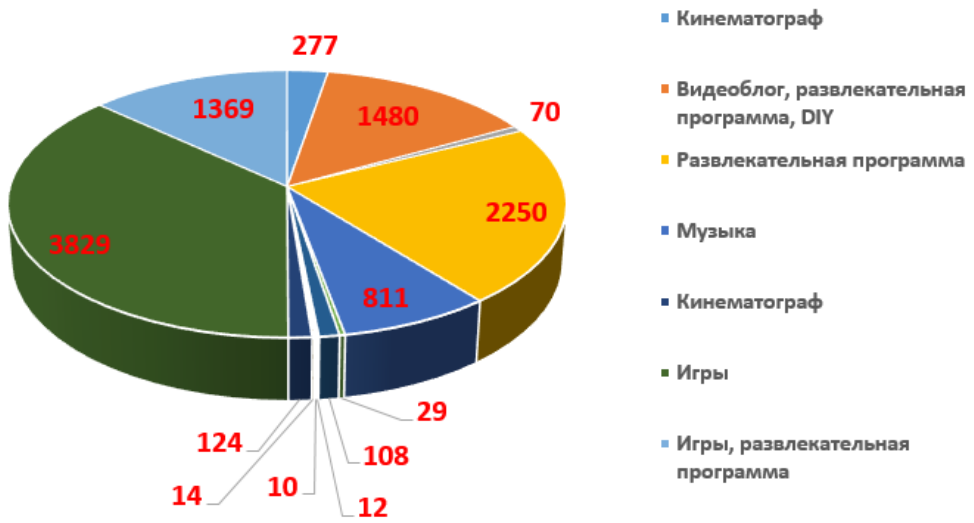

Рис. 15. Жанры видеороликов в сообществах русскоязычного сектора

Заключение

Целью работы было изучение структурных особенностей сети каналов *YouTube* – распределения степеней, характера корреляций степеней вершин, а также структуры сообществ. Для проведения исследований были собраны различные карты подсетей исследуемой сети. Результаты анализа показывают, что построенная сложная сеть каналов *YouTube* является мало-мировой, масштабно-инвариантной с показателем, равным 2, относится к слабо-дисассортативным (причем дисассортативность носит структурный характер) и имеет хорошо выраженную структуру сообществ, коррелирующую с главными жанрами видеоконтента. Была изучена динамика роста степеней узлов, показано, что сеть является растущей с нелинейным предпочтительным присоединением с показателем, равным 0.7.

Полученные результаты позволяют генерировать сети, статистически подобные сети каналов *YouTube*, и исследовать в них особенности распространения информации [9].

Список источников

1. Евин И. А. Введение в теорию сложных сетей // Компьютерные исследования и моделирование. – 2010. – Т. 2, № 2. – С. 121–141. DOI: https://doi.org/10.20537/2076-7633-2010-2-2-121-141
2. Complex Network // Википедия : свободная энциклопедия. – URL : https://en.wikipedia.org/wiki/Complex_network (дата обращения: 10.06.2021).
3. Что такое YouTube: история YouTube и интересные факты // ОПТИМИЗМ.РУ : настрой на результат. – URL: https://wiki.optimism.ru/Что_такое_YouTube:_история_YouTube_и_интересные_факты (дата обращения: 10.06.2021).
4. Paolillo J. C., Ghule S., Harper B. P. A Network View of Social Media Platform History: Social Structure, Dynamics and Content on YouTube // Proceedings of the 52nd Hawaii International Conference on System Sciences — 2019. — URL: http://hdl.handle.net/10125/59701 (дата обращения: 12.06.2021)
5. User and group networks on YouTube: A comparative analysis / M. Jebabli, H. Cherifi, S. Cherifi, A. Hamouda // 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA). – 2015. С. 1-8. – DOI : 10.1109/AICCSA.2015.7507126 (дата обращения: 12.06.2021).
6. YouTube API v3 // Google Develorers : [веб-сайт]. — URL: https://developers.google.com/youtube/v3 (дата обращения: 15.06.2021).

7. Network Data Repository. – Ryan Rossi; Nesreen Ahmed, 2012-2022. — URL: http://networkrepository.com/ (дата обращения: 15.06.2021).

8. Bernhard R. YouTube Data Tools. – URL: https://tools.digitalmethods.net/netvizz/youtube/index.php (дата обращения: 17.06.2021).

9. Barabási A.-L. Network Science. – URL: http://networksciencebook.com/chapter/10#introduction10 (дата обращения: 21.06.2021).

10. Clauset A., Shalizi C. R., Newman M. E. J. Power-law distributions in empirical data // SIAM Review. – 2009. – Т. 51, № 4. – С. 661-703. – DOI: 10.1137/070710111. – URL: https://arxiv.org/abs/0706.1062v2 (дата обращения: 25.06.2021).

11. Fast unfolding of communities in large networks / V. D.Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre // Journal of Statistical Mechanics: Theory and Experiment. – 2008. – Т. 2008, № 10. – С.P10008. – URL: https://arxiv.org/abs/0803.0476 (дата обращения: 30.06.2021).