ИССЛЕДОВАНИЕ ОСОБЕННОСТЕЙ РАСПРОСТРАНЕНИЯ ИНФОРМАЦИИ В СОЦИАЛЬНОЙ СЕТИ ВИДЕОХОСТИНГА YOUTUBE

Дидоренко А.В., Прогулова Т.Б.

В работе исследуется влияние структурных и топологических особенностей социальной сети видеохостинга YouTube на процессы распространения информации. Вычислены и проанализированы базовые характеристики сети и показатели центральности. Внимание фокусируется на топологических особенностях, включая структуры сообществ и ядро-периферия. В ходе исследования использовалась модель распространения, учитывающая выявленные свойства сети. Изучено влияние структуры сообществ на процессы распространения информации, а также исследована роль значимых узлов на масштаб и время распространения информации. Полученные результаты могут быть основой для решения задач поиска суперраспространителей, блокировки негативного влияния, формирования наборов наиболее влиятельных вершин для решения задач распространения и блокировки.

doi: 10.20537/mce2024econ05

Введение. Феномен распространения информации играет важнейшую роль, его изучение направлено на создание строгой аналитической и численной платформы для количественной оценки и прогнозирования распространения в различных областях [1]. В социальной сфере одной из мощных сред распространения информации является *YouTube*, за последние годы превратившийся в совершенный ресурс обмена разного рода информацией. Процессы на уровне *YouTube*-каналов протекают с учетом структуры сети подписок, не являющейся тривиальной, но под управлением системы рекомендаций, принцип работы которой в целом не известен.

Естественно предполагать, что для управления распространением, как и моделирования процессов распространения информации требуется понимание структуры исходной сети и определения наиболее влиятельных узлов [1]. Таким образом, целью работы было исследование влияния структурных и топологических особенностей социальной сети *YouTube*

на процессы распространения информации. Основные этапы работы: сбор данных для построения сети, анализ структуры сети и, наконец, изучение влияния свойства безмасштабности, структуры сообществ и влиятельных (центральных) узлов сети на процессы распространения.

Метод исследования и построение выборки. Видеохостинг *YouTube* включает в себя гигантскую социальную сеть, — ее исследование возможно только на уровне подсетей. Полная сеть *YouTube* доступна только *Google* / *YouTube*, а данные для видео доступны только через специальные функции. В данной работе анализируется сеть, узлам которой соответствовали *YouTube*-каналы, а направленным связям — отношения подписки между каналами (от подписчика к тому каналу, на который он подписан).

Для изучения влияния на процессы распространения структурных особенностей (в первую очередь, безмасштабности и структуры сообществ) исследование проводилось на трех сетях: (1) эмпирическая сеть, построенная на основании собранных данных, (2) рандомизированная эмпирическая сеть, полученная путем пересвязывания вершин с сохранением распределения степеней для ослабления структуры сообществ сети и (3) случайная сеть Эрдеша-Реньи с таким же количеством вершин и ребер, что и эмпирическая сеть, но с однородно случайным характером связей (принципиально отсутствие масштабно-инвариантного свойства). Для построенных сетей вычислялись основные глобальные и локальные характеристики, определялась структура сообществ и ядро-периферия.

В качестве эмпирической сети была выбрана сеть *YouTube*-каналов новостного характера, так как процессы распространения информации лучше всего изучать на том сегменте, где рассылка информации протекает активнее всего, а именно в СМИ. В настоящем исследовании сбор данных начинался от 100 наибольших по числу подписчиков новостных *YouTube*-каналов [3]. Процедура сбора данных описана в [4]. Отметим, что информация распространяется от канала к тем каналам, которые на него подписаны, поэтому для моделирования распространения информации от канала-подписки к подписчику список ребер был инвертирован.

С учетом выявленных особенностей эмпирической сети была определена модель распространения [2]. Изучалось влияние структуры сообществ на распространение, а также влияние выбора значимых вершин в качестве источников распространения на максимизацию охвата сети.

Анализ структуры сети каналов *YouTube*. Для построенной сети *YouTube*-каналов определялись распределение степеней, показатель нелинейности предпочтительного присоединения, характер корреляций

степеней вершин, коэффициент кластеризации, значения центральностей вершин и их распределения, структуры сообществ и ядро-периферия. Методика анализа структуры сети частично описана в работе [4].

Эмпирическая сеть обладает следующими характеристиками: количество узлов — 13 604, количество связей — 123 344, диаметр сети — 19, среднее расстояние — 5.6, а средняя полустепень захода/исхода равна 9.06. Таким образом, для сети новостных каналов в *YouTube* характерно небольшое среднее расстояние. Значение коэффициента кластеризации, на первый взгляд, не велико — 0.074, но коэффициент кластеризации для случайной сети Эрдеша-Реньи с тем же количеством вершин и ребер (или при рандомизации исходного графа) равен 0.0014 (почти в 53 раза меньше). Это говорит о существенной кластеризации сети каналов *YouTube*. Распределения степеней анализируемой сети показаны на рис.1.

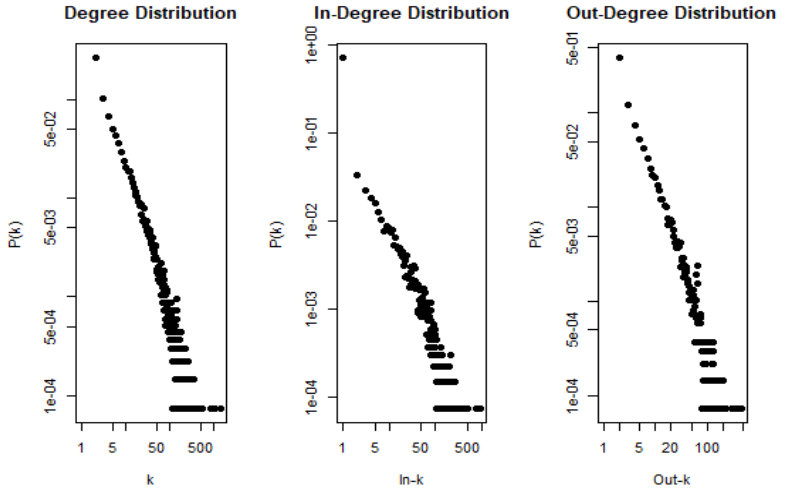

Рис. 1. Распределение степеней, полустепеней захода и исхода новостной сети в *log-log* масштабе.

Форма распределений (прямолинейность в *log-log* масштабе) указывает на то, что изучаемая сеть является масштабно-инвариантной. Для вычисления показателя распределение степеней изучаемой сети

аппроксимировалось степенной функцией $P(k) \sim k^{-\gamma}$ методом максимального правдоподобия. Результаты показали, что $\gamma \approx 2.01 \pm 0.02$.

Проводилась *k*-ядерная декомпозиция эмпирической сети для определения структуры ядро-периферия. На рис.2 показано распределение полученных значений показателя *k-shell* в *log-log* масштабе для ненаправленной сети. Из распределений видно, что подавляющее большинство узлов имеют небольшие значения *k-shell.* Максимальное значение *k-shell*, равное 64, имеют 66 вершин. Такая картина позволяет говорить о том, что эмпирическая сеть обладает структурой ядро-периферия, где ядрами являются вершины с высоким значением *k-shell*, а остальные вершины можно рассматривать как периферийные.

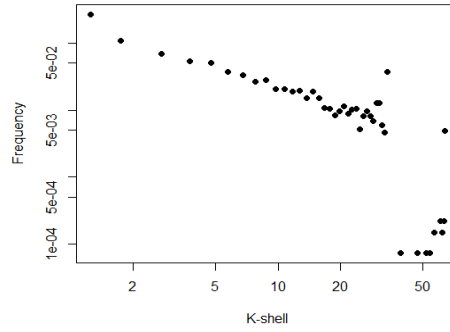

Рис. 2. Распределение значений *k-shell* эмпирической сети в *log-log* масштабе.

В рамках сложно-сетевого подхода предложен ряд показателей для ранжирования узлов по важности — так называемые центральности узлов. Основные такие показатели были вычислены для узлов всех трех сетей, на рис. 3 представлены их распределения в *log-log* масштабе. Графики указывают, что большая часть узлов имеет невысокие значения центральностей, количество с высокими значениями невелико. Можно предположить, что узлы с высокими значениями центральностей наиболее влиятельны. В рандомизированной сети с сохранением степеней некоторые при стирании структуры сообществ распределения центральностей меняются. Так, в эмпирической сети наблюдается несколько вершин с высоким значением *vertex betweenness (центральность вершины по посредничеству)*, а в рандомизированной сети всего одна вершина.

Распределения для центральности *PageRank* различаются по форме и диапазону значений. В случае случайной сети пуассоновский характер распределений говорит об однородности: все вершины имеют значения центральностей, близкие к среднему. Можно заключить, что при «стирании» структуры сообществ при рандомизации с сохранением степеней, появляются значимые различия в значениях таких центральностей, как *vertex betweenness, PageRank*, *eigenvector* (центральность по собственному вектору), *hub score* (центральность по информативности).

Также изучались корреляции значений центральностей. Корреляционные матрицы, отражающие коэффициенты корреляции Пирсона между всеми парами характеристик показаны на рис. 4. Почти все центральности и степени вершин положительно коррелируют между собой, что объясняет схожую форму их распределений. В эмпирической сети наблюдается слабая отрицательная корреляция между *closeness* (центральность по близости) и другими центральностями. При рандомизации сети с сохранением распределения степеней некоторые центральности начинают сильнее коррелировать между собой, например, значения центральности собственного вектора сильнее коррелируют со степенями вершин, *PageRank*, *authority* (центральность по влиятельности) и *hub score*. Возможно, структура сообществ ослабляет корреляцию между значениями центральностей узлов. Матрица корреляций значений центральностей узлов в случайной сети Эрдеша-Реньи демонстрирует отсутствие отрицательной корреляции.

Наконец, изучалась структура сообществ в сети *YouTube*-каналов. Для выделения сообществ использовался алгоритм *Louvain*, который демонстрирует высокую эффективность на больших графах [5]. Полученные результаты отражены на рис. 5, где показано значение модульности M, количество найденных сообществ, а также их размеры.

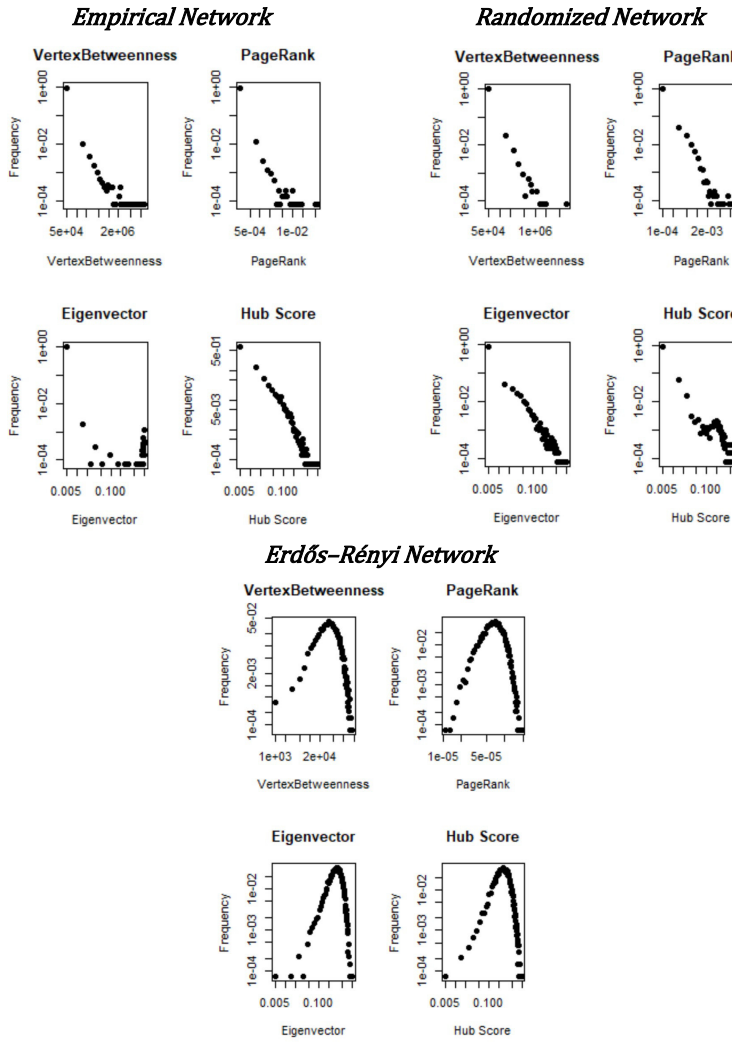

Рис. 3. Распределение центральностей для эмпирической сети, рандомизированной с сохранением степеней сети, случайной сети Эрдеша-Реньи.

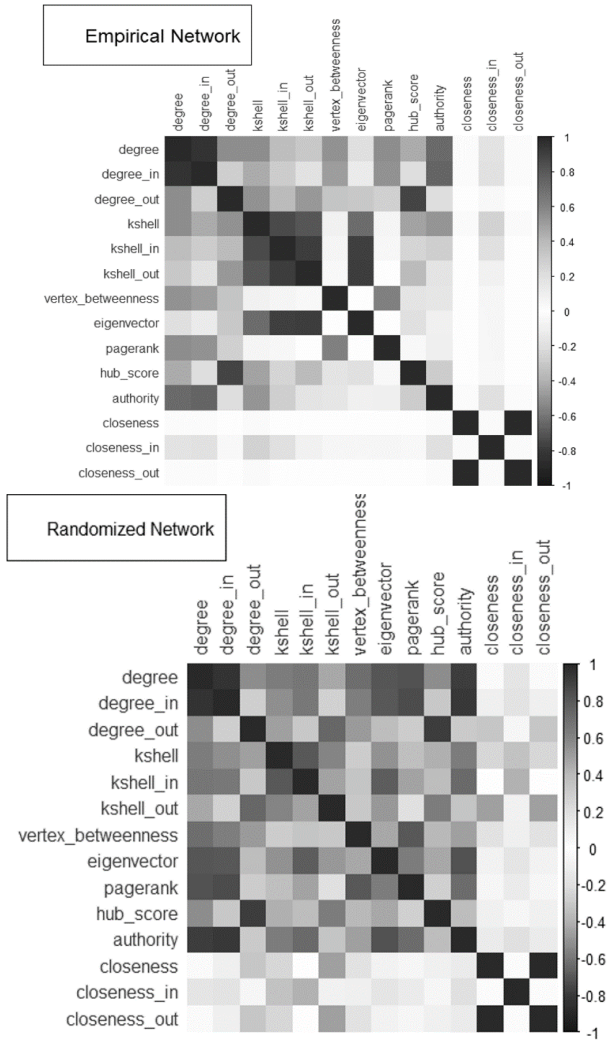

Рис. 4. Матрицы корреляций показателей центральности для эмпирической сети (сверху), рандомизированной с сохранением степеней сети (снизу).

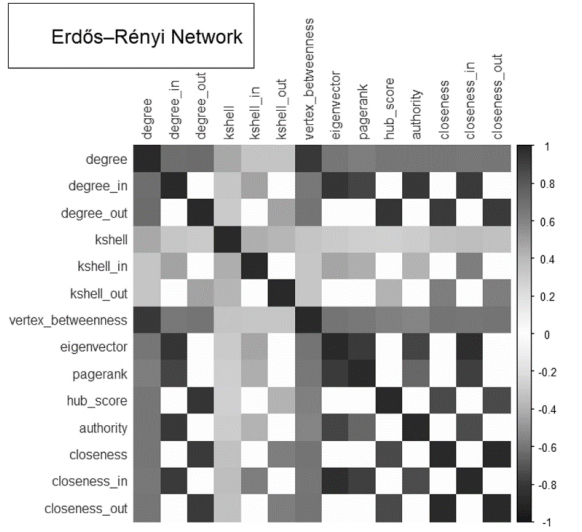

Рис. 4. Продолжение. Матрицы корреляций показателей центральности для случайной сети Эрдеша-Реньи**.**

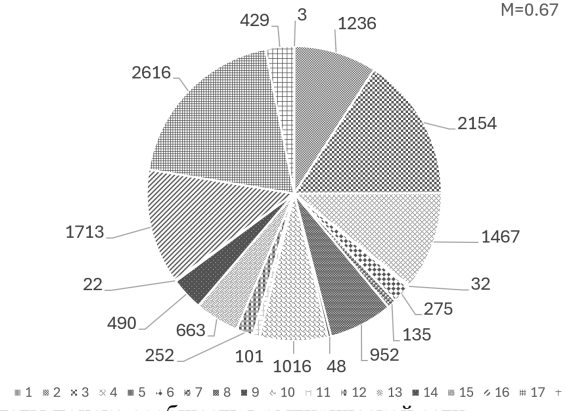

Рис. 5. Результаты поиска сообществ в эмпирической сети.

Исследование процессов распространения информации. Прежде всего, зафиксируем сценарий распространения информации, который в дальнейшем будет имитироваться. Канал, который опубликовал новый видеоролик, из которого пользователи могут получить новые идеи, отображается на главной странице *YouTube* для всех подписчиков. Затем подписчики могут создавать собственные видеоролики, обсуждая эти идеи, и их подписчики могут продолжить процесс распространения или проигнорировать идею. Процесс может продолжаться, пока будут встречаться каналы-подписчики, готовые передавать идею дальше.

Процессы распространения имитировались на эмпирической, рандомизированной с сохранением степеней и случайной сетях. На начальном этапе распространение информации имитировалось посредством модели нулевого уровня *SIR* [6]. Проводились компьютерные имитации процессов при произвольных значениях параметров β и γ, каждый запуск инициировался от случайно выбранной вершины сети. На рис.6 приведены усредненные результаты имитаций процессов от случайно выбранных вершин сетей для параметров $\beta = 0.8$ и $\gamma = 0.8$ для всех сетей. В случайной сети процессы затрагивают большую часть узлов, в то время как в эмпирической сети таких узлов меньше (см. рис. 6). Отличие связано с однородностью случайной сети, в то время как в эмпирической сети есть хабы и есть структура сообществ. Результаты для рандомизированной сети показывают, что процессы распространения охватывали большее число узлов за более короткое время по сравнению с эмпирической сетью (см. рис. 6). Это свидетельствует о значимости для процесса распространении заражения (информации) не только неоднородности (безмасштабности) сети, но и структуры сообществ.

Однако в модели *SIR* отсутствует возможность изучения непосредственного влияния на распространение информации структурных особенностей. Сообщества в *YouTube* в основном являются тематически обусловленными, естественно предполагать, что вероятность передачи (восприятия) информации между узлами одного сообщества больше, чем между узлами разных сообществ. Более адекватной в таком случае является иерархическая каскадная модель распространения на основе связей в сети — *Edge Based Hierarchical* (*EBH*) *Model*, учитывающая структуры сообществ и ядро-периферия [2]. Согласно *EBH* модели, в момент $t = 0$ начальные узлы являются активными, а остальные узлы не активными.

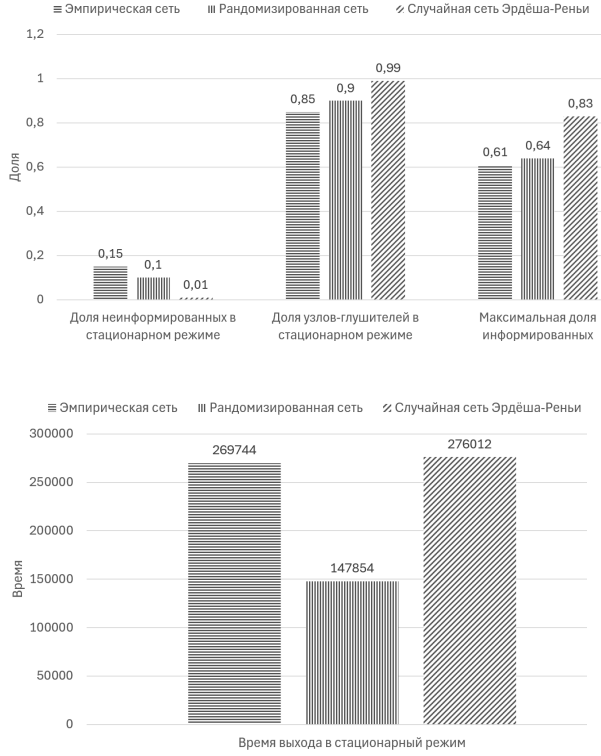

Рис. 6. Усредненные результаты имитаций процессов от случайно выбранных вершин сетей при $\beta = 0.8$ и $\gamma = 0.8$ в рамках модели *SIR*.

На шаге $t + 1$ каждый узел, активированный в предыдущий момент t, может активировать своих узлов-соседей в соответствии с иерархией вероятностей, определяемой классом ребра:

$$p_{cc} > p_{cp} > p_{pp_0} > p_{pp_1} > p_{pc}. \quad (1)$$

Распространение заканчивается в то время, когда не было активировано ни одного нового узла [2]. Вероятность распространения привязана к каждому ребру E_{ij}, класс ребра зависит от типов обоих смежных узлов, i и j. В модели *EBH* ребра разделены на пять классов:

$$\begin{cases} E_{cc} = \{E_{ij} \in E : i \in Core \text{ and } j \in Core\} \\ E_{cp} = \{E_{ij} \in E : i \in Core \text{ and } j \in Per\} \\ E_{pc} = \{E_{ij} \in E : i \in Per \text{ and } j \in Core\} \\ E_{pc} = \{E_{ij} \in E : i \in Per \text{ and } j \in Core\} \\ \begin{cases} E_{pp_0} = \{E_{ij} \in E : \delta_{ij} = 0\} \\ E_{pp_1} = \{E_{ij} \in E : \delta_{ij} = 1\} \end{cases} \\ Core = \{x | x - \text{узел с максимальным значением } k_s \; Per = V \backslash Core\} \end{cases} \quad (2)$$

Индекс «*c*» означает принадлежность узла ядру сети, а «*p*» — периферии. Процессы распространения изучались в рамках *EBH*-модели на эмпирической, рандомизированной с сохранением степеней и случайной сетях. В эмпирической и рандомизированной сетях каждому ребру были установлены вероятности передачи информации по ребрам, исходя из класса ребра (3) в иерархии вероятностей (2). Для эмпирической и рандомизированной сетей здесь зафиксирована следующая произвольная иерархия вероятностей для ребер: $0.9 > 0.7 > 0.5 > 0.3 > 0.1$. Применить аналогичную иерархию для случайной сети не предоставляется возможным ввиду отсутствия в ней структуры сообществ. Поэтому для случайной сети была взята вероятность 0.454, средняя по всем ребрам эмпирической сети. На рис.7 приведены усредненные результаты моделирования процессов от случайных вершин сетей в рамках модели *EBH* для всех сетей.

В случайной сети процессы охватили почти все узлы сети, в эмпирической — меньше половины узлов, в рандомизированной сети информационный охват больше, чем в эмпирической (см. рис. 7). Как и в случае модели *SIR*, такие результаты позволяют сделать вывод, что структура сообществ влияет на охват узлов, уменьшая его.

Для проверки гипотезы о максимизации охвата и минимизации времени распространения при инициации процесса от наиболее значимых узлов проводилась серия экспериментов, в которых процесс запускался от вершин с максимальными значениями центральностей, а также от случайных вершин первых трех оболочек с максимальными значениями *k-shell*. Симуляции проводились на всех сетях в рамках моделей *SIR* и *EBH* (см. рис. 8). Результаты симуляций показывают, что запуск от центральных вершин или вершин из разных оболочек не приводит к значительно большему охвату узлов, чем при случайном выборе вершины (см. рис. 6–7).

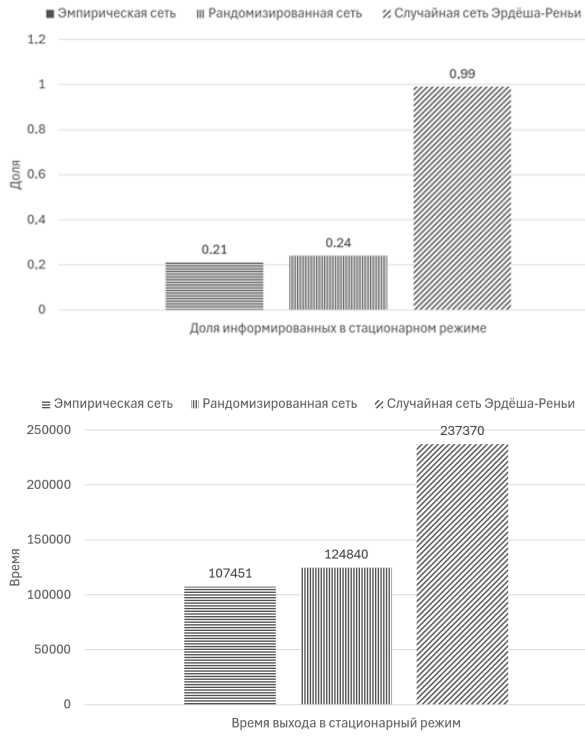

Рис. 7. Усредненные результаты моделирования процессов от случайных вершин сетей в рамках модели *EBH.*

Однако, моделирование в рамках *EBH*-модели показало, что запуск от некоторых центральных вершин приводит к значительному сокращению времени. Это может быть полезно, например, при распространении срочной информации.

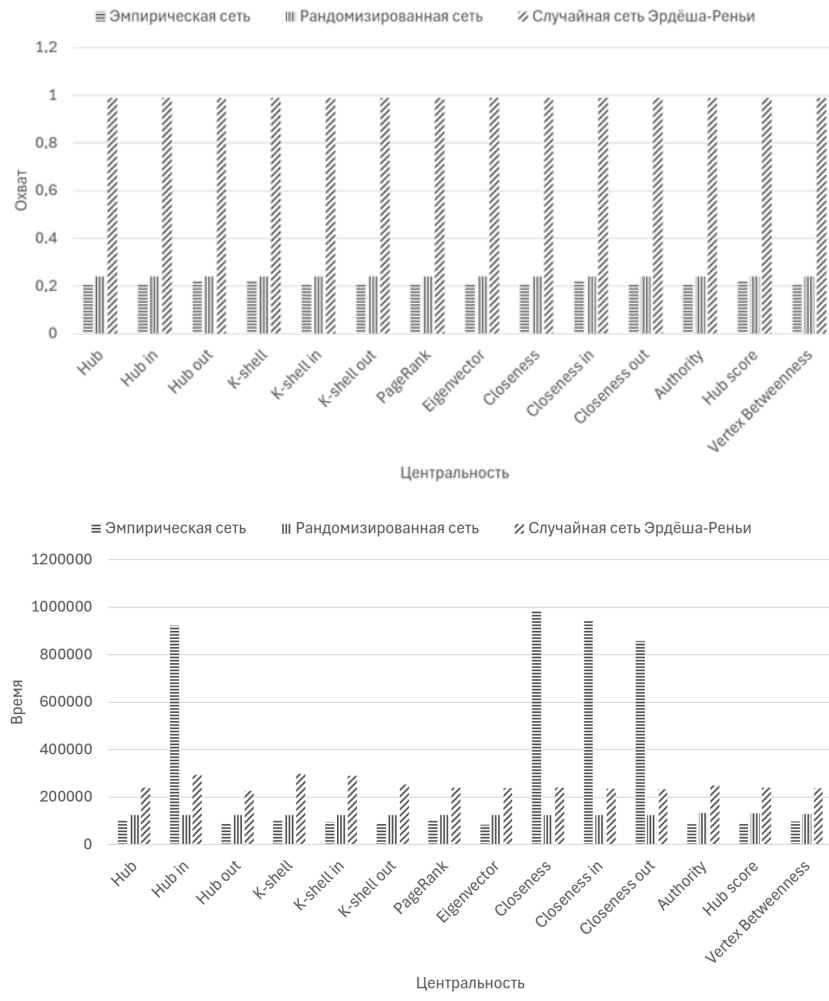

Рис. 8а. Результаты экспериментов (показатели времени и охвата) в случае старта от центральных вершин для модели EBH.

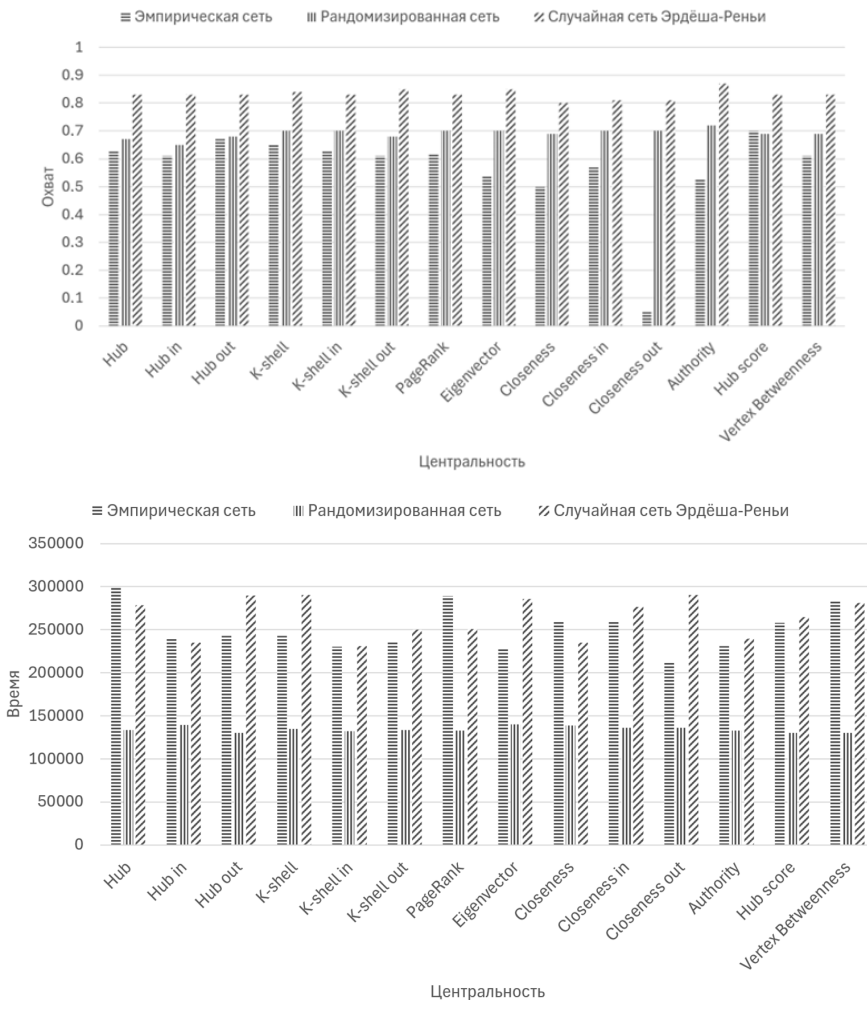

Рис. 8б. Результаты экспериментов (показатели времени и охвата) в случае старта от центральных вершин для модели SIR.

Заключение. На основании проведенного исследования можно утверждать, что сеть каналов *YouTube* является безмасштабной и имеет хорошо выраженные структуры сообществ и ядро-периферия. Изучение процессов распространения показало, что запуск распространения от центральных вершин не максимизирует информационный охват сети, но для показателей центральности, как например, *PageRank*, *eigenvector* и *vertex betweenness* время процесса на порядок ниже. Наконец, при ослаблении структуры сообществ масштаб распространения увеличивается, что позволяет рассматривать сообщества как «ловушки» для распространения информации. В случае же полного уничтожения структуры можно наблюдать почти полный охват сети.

Полученные результаты могут быть основой для решения задач поиска суперраспространителей, создания эффективных стратегий блокировки негативного влияния, формирования наборов наиболее влиятельных вершин для распространения и блокировки.

Отметим, что полученные результаты являются оценочными, так как для более точного и глубокого исследования требуется больше статистических данных и модельных экспериментов.

СПИСОК ЛИТЕРАТУРЫ

1. *Barabasi, A.-L.* Network Science / A.-L. Barabasi. – Cambridge: Cambridge University Press, 2016.
2. *Gupta, Y.* Dynamics of Information Diffusion on Online Social Networks / Y. Gupta. – Rupnagar: Indian Institute of Technology Ropar, 2017.
3. Socialblade: [project] // List of most-subscribed YouTube channels: [web platform]. 2023. URL: https://socialblade.com/youtube/top/category/ news
4. *Дидоренко, А. В., Прогулова Т. Б.* Построение и исследование структуры сложной сети *YouTube* –каналов // *Системный анализ в науке и образовании: сетевое научное издание*. 2022. *No* 1. С. 77-90. *URL*: http://sanse.ru/download/462.
5. *Blondel, V.D.* Fast unfolding of communities in large networks / V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre // arXiv.org : [open archive]. – 2008. – arXiv: 0803.0476 [gr-qc]. URL: https://arxiv.org/abs/0803.0476.
6. *Kermack, W.O., McKendrick, A.G.* A contribution to the mathematical theory of epidemics // *Proc. R. Soc. Lond. Ser. A*. 1927. Vol.115(772). P.700–721.

STUDYING THE FEATURES OF INFORMATION PROPAGATION PROCESSES IN YOUTUBE VIDEO HOSTING SOCIAL NETWORK

Didorenko A.V., Progulova T.B.

The paper explores the influence of structural and topological features of the social network of YouTube video hosting on the processes of information propagation. Basic network characteristics and centrality measures are computed and analyzed. Attention is focused on topological features, including community structures and core-periphery. A propagation model that takes into account the identified network properties is used in the study. The influence of community structure on information propagation processes is studied, and the role of significant nodes on the scale and time of information propagation is investigated. The obtained results can be a basis for solving the problems of searching for superpropagators, blocking negative influence, forming sets of the most influential nodes to solve the problems of propagation and blocking.