ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 24 за 2021 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Ю.Н. Орлов, А.С. Панкратов

К разработке модели эволюции структуры сетевого графа

Рекомендуемая форма библиографической ссылки: Орлов Ю.Н., Панкратов А.С. К разработке модели эволюции структуры сетевого графа // Препринты ИПМ им. М.В.Келдыша. 2021. № 24. 16 с. https://doi.org/10.20948/prepr-2021-24
https://library.keldysh.ru/preprint.asp?id=2021-24

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

Ю.Н. Орлов, А.С. Панкратов

К разработке модели эволюции структуры сетевого графа

Москва — 2021

Орлов Ю.Н., Панкратов А.С.

К разработке модели эволюции структуры сетевого графа

В работе предложена модель исследования структуры сетевого графа с многомерным распределением вершин по степеням. В качестве примера характерной структуры рассмотрен граф, образованный дружескими связями социальной сети «ВКонтакте». Рассмотрено распределение степеней вершин графа по набору параметров, характеризующих пользователей. Для моделирования эволюции структуры графа применено демографическое уравнение и уравнение типа Лиувилля.

Ключевые слова: сетевой граф, эволюция распределения степеней вершин, распределение рангов, многомерные вершины

Orlov Yu.N., Pankratov A.S.

To the evolution model of network graph structure construction

In this paper the investigation of the structure of network graph is presented. The social network between the Russian towns is considered. It is shown, that the distribution of vertex powers is uniform. As a consequence there is a high dimension region with whole connection. The probability of special sub-graphs is estimated. The Liouville equation is used for modeling of the graph structure evolution.

Key words: network graph, rang distribution, power vertex evolution, multidimensional graph

Работа выполнена при поддержке гранта РФФИ № 19-01-00602.

Содержание

1. Введение и постановка задачи

Анализ графов социальных сетей представляет собой весьма популярную прикладную математическую задачу, имеющую также и важный теоретический аспект, связанный с разработкой алгоритмов быстрого поиска определенных конфигураций. Теоретическая сторона задачи интересна в первую очередь тем, что в настоящее время отсутствуют доказательные утверждения относительно статистических свойств больших связных графов, позволяющие строить эффективные алгоритмы поиска. Большая часть таких алгоритмов связана с модельными графами (типа «колесо», «звезда», «пропеллер», «дерево» и т.п.) либо основана на эвристиках, проверяемых лишь постфактум. В качестве примеров теоретических работ, в которых обсуждаются основные характеристики графов социальных сетей и алгоритмы, их вычисляющие, следует указать работу [1], в которой обсуждаются некоторые практические методы анализа, работу [2] об алгоритмах визуализации и анализа графов больших размеров, а также работу [3] с подробным анализом методов исследования сетевых структур.

Актуальной прикладной задачей является разработка алгоритмов моделирования сетевой структуры графа с определенной функцией распределения вершин по степеням, а также оценки количества путей с определенными свойствами, например, циклов заданной длины. В настоящее время анализ графов носит либо конкретный прикладной характер, связанный с приближенным моделированием реального графа подходящим случайным типа Эрдёша-Реньи [4], либо изучаются свойства графов, построенных по определенному теоретическому правилу. В обоих случаях рассматривается статичная структура графа, что не позволяет оценить влияние внешних воздействий на такую систему. Плотность распределения степеней вершин реальных графов социальных сетей хорошо аппроксимируется степенной зависимостью [5, 6], так что основным направлением моделирования является построение случайных графов с таким распределением. При этом под эволюцией графа понимается изменение его структуры при добавлении вершин и ребер в процессе работы соответствующего алгоритма. Однако такой подход не позволяет анализировать изменение структуры связей при определенном воздействии на информационную среду, то есть не позволяет анализировать ситуацию с меняющимся по времени распределением степеней вершин. Этот аспект представляется важным, поскольку, например, структура связей между пользователями с определенными возрастами уже сама по себе не является стационарной в силу меняющегося возраста абонентов.

Подходы к анализу структуры социальных сетей описаны в [5-8] и представляют собой способы аппроксимации наблюдаемого распределения степеней вершин подходящим случайным графом. В более общей постановке проблемой является исследование нестационарных взаимосвязей между статистиками степеней многомерных вершин сетевых графов. Такая задача

возникает при изучении связей между пользователями социальных сетей, когда наряду с отдельной вершиной графа, представляющей собой (предположительно однозначно) индивидуального пользователя, возникает необходимость исследовать распределение связей между классами пользователей, объединенных по совпадающим признакам в соответствии с анкетированием: пол, возраст, место проживания и/или рождения, образование, область интересов и т.п.

В настоящей работе исследуется структура связей неориентированного графа, представляющего социальную сеть «дружбы» между пользователями определенного пола и возраста, проживающими в регионах России – крупных городах и областях. Данные для анализа взяты с ресурса «ВКонтакте» за 2015 г. в виде, обработанном в [9] и описанном в [10-12]. Рассматривается трехпараметрическая структура вершин графа (пол, возраст, город), где каждый параметр имеет свою размерность. Эти параметры естественно считать независимыми, поэтому выявление статистически значимых групп связей между ними важно для разработки модели такого графа.

2. Многомерные распределения степеней вершин

Пусть каждая вершина на уровне отдельного пользователя определяется набором параметров $x_1, x_2, ..., x_s$, которые обозначают различные анкетируемые параметры – пол, возраст, национальность, регион проживания, уровень образования, область интересов и т.д. Совокупность характеристик всех пользователей сети в данный момент времени образует s-мерное дискретное пространство параметров, шаг дискретизации определяется масштабом инструмента измерения. Каждому пользователю отвечает точка в этом пространстве, связям между ними – отрезок без направления. Размерность совокупности характеристик пользователей по каждому из трех параметров обозначим соответственно $X_1, X_2, ..., X_s$. Например, для параметра «пол» естественно ввести три варианта ответов ($X_1 = 3$): «мужской», «женский» и «не указан», для параметра возраст – примерно сто вариантов по годам в соответствии с годом рождения и т.д. Следует подчеркнуть, что мы не обсуждаем соответствие параметров вершин действительным характеристикам пользователей, они могут позиционироваться в сети как угодно.

Введем число связей между пользователями, которые объединены в заданные классы: $N_{i_1 \ldots i_s}^{j_1 \ldots j_s}$ есть число связей между пользователями классов, которым отвечает мультииндекс i, с пользователями классов j. По парам (i_k, j_k) верхних и нижних индексов число связей $N_{i_1 \ldots i_s}^{j_1 \ldots j_s}$ симметрично. Эта $2s$-индексная величина представляет основную исходную характеристику изучаемой системы. Заметим, что поскольку такое изучение в контексте больших данных основано на сборе статистической информации о разных пользователях, надо оценить количество состояний системы, т.е. потенциальное

число ячеек, требующих заполнения. Однако уже количество ячеек для трех самоочевидных параметров пол-возраст-регион оценивается в [12] величиной порядка 10^{11}, что примерно в 10 раз больше всего населения планеты. Поэтому получить данные о структуре таких многопараметрических сетей статистическим путем невозможно: не хватит пользователей, чтобы достичь приемлемой репрезентативности выборки, не говоря уже о том, чтобы получать статистически достоверные выводы о социальных связях между группами по уровню образования, национальности, социальному статусу и т.д. Следовательно, необходимо применить сокращение описания посредством рассмотрения кластеризации вершин исходного графа по набору параметров.

Удобно ввести обозначение $S^{(i)}_{(j)} N^{j_1 \dots j_s}_{i_1 \dots i_s}$ для описания результата суммирования числа связей по заданному подмножеству верхних и нижних индексов. Например, $S^{i_k} N^{j_1 \dots j_s}_{i_1 \dots i_s}$ означает суммирование связей по пользователям с характеристикой i_k, величина $S^{i_k}_{j_k} N^{j_1 \dots j_s}_{i_1 \dots i_s}$ показывает структуру графа безотносительно k-ой характеристики и т.д. В этом смысле (но не в тензорном) операцию $S^{(i)}_{(j)} N^{j_1 \dots j_s}_{i_1 \dots i_s}$ можно назвать «сверткой» многомерного распределения связей между параметрами абонентов сети.

Обозначим также

$$ {}_k f^{j_k}_{i_k} = S^{(i_1 \dots \hat{i}_k \dots i_s)}_{(j_1 \dots \hat{j}_k \dots j_s)} N^{j_1 \dots j_s}_{i_1 \dots i_s} \qquad (1) $$

результат свертки по всем парам индексов, кроме k-ой пары. Получающийся объект – своеобразная «проекция» многомерного графа на пространство k-го параметра. Например, можно просуммировать связи по всем параметрам, кроме «пола», и получить в результате взвешенный граф в виде треугольника с циклическими связями у каждой вершины. Обозначим матрицу смежности для распределения связей как ${}_k F_i{}^j$. Ее элементы равны нулю, если ${}_k f_i{}^j = 0$, и единице, если ${}_k f_i{}^j > 0$.

Если требуется проанализировать распределение связей между пользователями с заданной (k-ой) характеристикой при меняющейся m-ой характеристике, то такое распределение получается при суммировании вида

$$ {}_k f^{j_k}_{i_k}(m) = S^{(i_1 \dots \hat{i}_k \dots \hat{i}_m \dots i_s)}_{(j_1 \dots \hat{j}_k \dots j_s)} N^{j_1 \dots j_s}_{i_1 \dots i_s} = S^{(i_1 \dots \hat{i}_k \dots i_s)}_{(j_1 \dots \hat{j}_k \dots \hat{j}_m \dots j_s)} N^{j_1 \dots j_s}_{i_1 \dots i_s} . \qquad (2) $$

Например, если k – возраст, а m – пол, то выражение (2) определяет число связей между заданной парой возрастов i и j для пола m. Очевидно, для любых пар k и m должно выполняться условие

$$ {}_k f_i{}^j = \sum_m {}_k f_i{}^j(m) . \qquad (3) $$

Матрицы связей в (3) получаются суммированием из более общего объекта:

$$ {}_k f^{j_k}_{i_k}(m) \equiv S^{i_m}{}_{k,m} f^{j_k j_m}_{i_k i_m} , \quad {}_k f^{j_k}_{i_k} = S^{i_m}_{j_m}{}_{k,m} f^{j_k j_m}_{i_k i_m} . \qquad (4) $$

Соотношения вида (4) являются основными при анализе совместных распределений связей между многомерными параметрами пользователей.

Если область значений параметра k в (4) разбита на n_k классов, то количество пар в распределении ${}_k f_i^{\,j}(m)$ для каждого m равно $n_k(n_k+1)/2$. Это относительно небольшое число по сравнению с количеством пользователей, поэтому для каждой такой «проекции» достаточно данных для описания структуры графа. Следовательно, если бы в идеале все параметры были независимы, то распределение $N_{i_1..i_s}^{j_1\cdots j_s}$ представлялось бы в виде произведения $\prod_k {}_k f_i^{\,j}$, что позволило бы полностью описать связи графа по всем параметрам. Однако независимыми можно приближенно считать только три параметра – пол абонента, его возраст и город проживания. Строго говоря, эти параметры тоже зависимы – хотя бы потому, что вероятность иметь больший возраст больше у женщин, чем у мужчин, а также зависит от экологии рабочих мест в регионе. Далее мы обсудим, насколько корректно представление числа связей $N_{ijk}^{\alpha\beta\gamma}$ абонента пола i возраста j из города k с абонентом пола α возраста β из города γ в виде произведения однопараметрических распределений $g_i^{\alpha}, f_j^{\beta}, h_k^{\gamma}$. Соответствующие матрицы смежности будем обозначать $G_i^{\alpha}, F_j^{\beta}, H_k^{\gamma}$.

Для изучения взаимосвязей между параметрами сети следует изучить двухпараметрическое распределение связей – по полу и возрасту, по полу и городу, по возрасту и городу. Такие распределения получаются суммированием числа связей по одной паре соответствующих индексов. Обозначим эти распределения соответственно $b_{ij}^{\alpha\beta}, c_{ik}^{\alpha\gamma}, d_{jk}^{\beta\gamma}$. Чтобы построить упрощенную модель исходного семейства связей, рассмотрим сумму каждого из этих распределений по одному из индексов. Например, $\varphi_i^{\alpha}(j)=\sum_{\beta} b_{ij}^{\alpha\beta}$ есть число связей между полами для данного возраста j, а матрица $\psi_j^{\beta}(i)=\sum_{\alpha} b_{ij}^{\alpha\beta}$ есть число связей между возрастами для данного пола i. При этом выполняется нормировка на суммарные величины связей между полами и возрастами:

$$\sum_j \varphi_i^{\alpha}(j)=g_i^{\alpha}, \quad \sum_i \psi_j^{\beta}(i)=f_j^{\beta}. \tag{5}$$

Аналогично пусть $\lambda_k^{\gamma}(i)=\sum_{\alpha} c_{ik}^{\alpha\gamma}$ есть число связей между городами для пола i, а $\mu_i^{\alpha}(k)=\sum_{\gamma} c_{ik}^{\alpha\gamma}$ есть число связей между полами для города k, так что

$$\sum_i \lambda_k^{\gamma}(i)=\omega_k^{\gamma}, \quad \sum_k \mu_i^{\alpha}(k)=g_i^{\alpha}. \tag{6}$$

Далее, если ввести количества связей между возрастами для данного города k в виде $q_j^{\beta}(k) = \sum_{\gamma} d_{jk}^{\beta\gamma}$ и количество связей между городами для данного возраста j в виде $p_k^{\gamma}(j) = \sum_{\beta} d_{jk}^{\beta\gamma}$, то получим третье условие взаимности:

$$\sum_k q_j^{\beta}(k) = f_j^{\beta}, \quad \sum_j p_k^{\gamma}(j) = \omega_k^{\gamma}. \tag{7}$$

Соответствующие матрицы смежности для зависимостей типа $\varphi_i^{\alpha}(j)$ обозначим через $\Phi_i^{\alpha}(j)$ и далее: $\Psi_j^{\beta}(i), \Lambda_k^{\gamma}(i), M_i^{\alpha}(k), Q_j^{\beta}(k), P_k^{\gamma}(j)$. В силу соотношений (5-7) их элементы являются подмножествами элементов матриц смежности $G_i^{\alpha}, F_j^{\beta}, H_k^{\gamma}$, так что удобной мерой различий между этими матрицами является норма Фробениуса. Поскольку элементы матриц смежности равны только нулям или единицам, то квадраты элементов равны самим элементам. В таком случае норма Фробениуса имеет вид (например, для матрицы $\Phi_i^{\alpha}(j)$):

$$\left\| G_i^{\alpha} - \Phi_i^{\alpha}(j) \right\|_F = \sqrt{2}\sqrt{K_G - K_{\Phi}(j)}, \tag{8}$$

где через $K_{G,\Phi}$ обозначено соответственно число ребер в матрицах смежности. Чтобы сравнивать между собой близость между матрицами смежности для разных характеристик пользователей, удобно нормировать расстояние (8) на величину $N_G\sqrt{\rho_G}$ матрицы смежности G_i^{α}, где $\rho = 2K/N^2$ есть плотность графа. Тогда норма близости между матрицами смежности будет выражаться числом, заключенным между нулем и единицей:

$$\left\| G_i^{\alpha} - \Phi_i^{\alpha}(j) \right\| = \frac{\left\| G_i^{\alpha} - \Phi_i^{\alpha}(j) \right\|_F}{N_G\sqrt{\rho_G}} = \sqrt{1 - K_{\Phi}(j)/K_G}. \tag{9}$$

Аналогично рассматриваются нормы близости матриц смежностей для остальных параметров пользователей.

3. Моделирование эволюции числа связей между вершинами

При анализе изменения матрицы смежности графа с течением времени надо уметь сравнивать меняющиеся структуры связей. Поскольку одним из параметров является возраст абонентов, то на следующий год структура связей – во всяком случае в части возраста – заведомо изменится. Чтобы корректно провести анализ возрастной структуры сети, необходимо учесть движение по возрастной характеристике, то есть скорректировать изменение наблюдаемых связей с учетом демографического уравнения. При краткосрочном анализе можно считать коэффициенты физической смертности в единицу времени (год) постоянными, равными средним по популяции в данном регионе в зависимости

от пола и возраста. Обозначим $m_{ik}(j)$ вероятность умереть человеку возраста j лет пола i в регионе k. Тогда изменение распределения $R_{ik}(j,t)$ населения по возрастам в данном регионе при сдвиге по времени t на один шаг описывается уравнением:

$$R_{ik}(j,t+1) = \big(1 - q_{ik}(j-1)\big) R_{ik}(j-1,t). \tag{10}$$

Граничное условие воспроизводства населения, т.е. интегральное уравнение рождаемости для численности населения возраста 0 лет при краткосрочном анализе не используется, поскольку младенцы в социальных сетях физически не присутствуют.

Если теперь предположить, что демографические характеристики одинаковы как для абонентов социальных сетей, так и для остального населения региона, то в силу естественного хода событий изменение распределения парных связей между абонентами будет описываться уравнением:

$$N_{ijk}^{\alpha\beta\gamma}(t+1) = N_{i(j-1)k}^{\alpha(\beta-1)\gamma}(t) - m_{ik}(j-1)\sum_{\lambda\mu\nu} N_{i(j-1)k}^{\lambda\mu\nu}(t) - m_{\alpha\gamma}(\beta-1)\sum_{rsm} N_{rsm}^{\alpha(\beta-1)\gamma}(t)$$
$$+ \frac{1}{N(t)} m_{ik}(j-1) m_{\alpha\gamma}(\beta-1)\sum_{\lambda\mu\nu} N_{i(j-1)k}^{\lambda\mu\nu}(t)\sum_{rsm} N_{rsm}^{\alpha(\beta-1)\gamma}(t). \tag{11}$$

Отличие фактического распределения вершин по степеням от предполагаемого, которое дается формулой (11), позволит оценить источниковые члены и вероятности перехода. Источник $S_{ijk}^{\alpha\beta\gamma}(t)$ описывает в год t появление пар новых связей или исчезновение старых, а переходные вероятности $P_{ijk\to i'j'k'}^{\alpha\beta\gamma\to\alpha'\beta'\gamma'}(t)$ описывают эффект «переобувания», т.е. изменения профиля пользователя. В итоге наблюдаемое распределение вершин по степеням будет формироваться на основе следующего уравнения:

$$N_{ijk}^{\alpha\beta\gamma}(t+1) = N_{i(j-1)k}^{\alpha(\beta-1)\gamma}(t) - m_{ik}(j-1)\sum_{\lambda\mu\nu} N_{i(j-1)k}^{\lambda\mu\nu}(t) - m_{\alpha\gamma}(\beta-1)\sum_{rsm} N_{rsm}^{\alpha(\beta-1)\gamma}(t)$$
$$+ \frac{1}{N(t)} m_{ik}(j-1) m_{\alpha\gamma}(\beta-1)\sum_{\lambda\mu\nu} N_{i(j-1)k}^{\lambda\mu\nu}(t)\sum_{rsm} N_{rsm}^{\alpha(\beta-1)\gamma}(t) + S_{ijk}^{\alpha\beta\gamma}(t) + \tag{12}$$
$$+ \sum_{\substack{\alpha'\beta'\gamma' \\ i'j'k'}}{}^{*} P_{i'j'k'\to ijk}^{\alpha'\beta'\gamma'\to\alpha\beta\gamma}(t) N_{i'j'k'}^{\alpha'\beta'\gamma'}(t) - N_{i(j-1)k}^{\alpha(\beta-1)\gamma}(t) \sum_{\substack{\alpha'\beta'\gamma' \\ i'j'k'}}{}^{*} P_{i(j-1)k\to i'j'k'}^{\alpha(\beta-1)\gamma\to\alpha'\beta'\gamma'}(t).$$

Здесь звездочкой * обозначено для краткости условие отсутствия «переходов» между совпадающими состояниями.

Суммирование выражения (12) по тем или иным парам индексов даст уравнение относительно эволюции распределения, зависящего от меньшего числа параметров, что позволит определить корректность гипотезы факторизации применительно к эволюционной задаче. Рассмотрим, в частности, эволюцию распределения степеней вершин возрастных связей, т.е.

величину $f_j^{\beta}(t) = \sum_{ik\alpha\gamma} N_{ijk}^{\alpha\beta\gamma}(t)$. Введем $M_{ijk}(t) = S_{\alpha\beta\gamma} N_{ijk}^{\alpha\beta\gamma}(t)$ в соответствии с обозначением (1). Тогда из (12) получаем:

$$f_j^{\beta}(t+1) = f_{j-1}^{\beta-1}(t) - \sum_{ik} q_{ik}(j-1) N_{i(j-1)k}(t) - \sum_{\alpha\gamma} q_{\alpha\gamma}(\beta-1) N_{\alpha(\beta-1)\gamma}(t) +$$

$$+ \frac{1}{N(t)} \sum_{ik\alpha\gamma} q_{ik}(j-1) q_{\alpha\gamma}(\beta-1) N_{i(j-1)k}(t) N_{\alpha(\beta-1)\gamma}(t) + S_j^{\beta}(t) + \qquad (13)$$

$$+ \sum_{\substack{\alpha'\beta'\gamma' \\ i'j'k'}}^{*} \sum_{ik\alpha\gamma} P_{i'j'k' \to ijk}^{\alpha'\beta'\gamma' \to \alpha\beta\gamma}(t) n_{i'j'k'}^{\alpha'\beta'\gamma'}(t) - f_{j-1}^{\beta-1}(t) \sum_{\substack{\alpha'\beta'\gamma' \\ i'j'k'}}^{*} P_{i(j-1)k \to i'j'k'}^{\alpha(\beta-1)\gamma \to \alpha'\beta'\gamma'}(t).$$

Из (13) следует, что факторизация корректна либо при отсутствии переходов между другими стратами, когда переходные вероятности равны нулю, либо в случае факторизации переходных вероятностей. Последнее предположение не отвечает реальности, потому что смена имиджа во многом определяется как раз атрибутами абонента, так что факторизационная модель характерна для статичного рассмотрения картины связей графа. При изучении же связей графа в динамике требуется рассматривать полный мультииндексный объект, что невозможно по причине отсутствия данных необходимой полноты, что было отмечено выше. Поэтому анализ эволюции графов социальных сетей может быть проведен лишь качественно – например, путем сравнения близости матриц смежности. Далее мы кратко опишем стационарный срез сети «Вконтакте» за 2015 год.

4. Анализ сети «ВКонтакте»

Матрица смежности H_k^{γ} связей между городами применительно к данной системе была исследована в [12]. Выяснилось, что «городской» граф связный. Число ребер равно $K_H = 1\,876\,564$, средняя плотность $\rho_H = \frac{2K_H}{R(R-1)} \approx 0{,}63$, где $R = 2441$ (число городов или регионов). Эта плотность равна вероятности того, что произвольные две вершины «городского» графа являются связными, то есть того, что в данных двух городах существует хотя бы по одному пользователю произвольного пола и возраста, которые дружат между собой.

При анализе статистики степеней вершин графа с матрицей H_k^{γ} представляет интерес зависимость степени вершины от ее ранга при упорядочении вершин по убыванию степеней и собственно распределение вершин по рангам (рис. 1). Они слабо нелинейные, отличие от прямой линии заметно лишь в семи первых и примерно в ста пятидесяти последних рангах. Суммарное количество городов в указанных фрагментах равно соответственно 55 и 180.

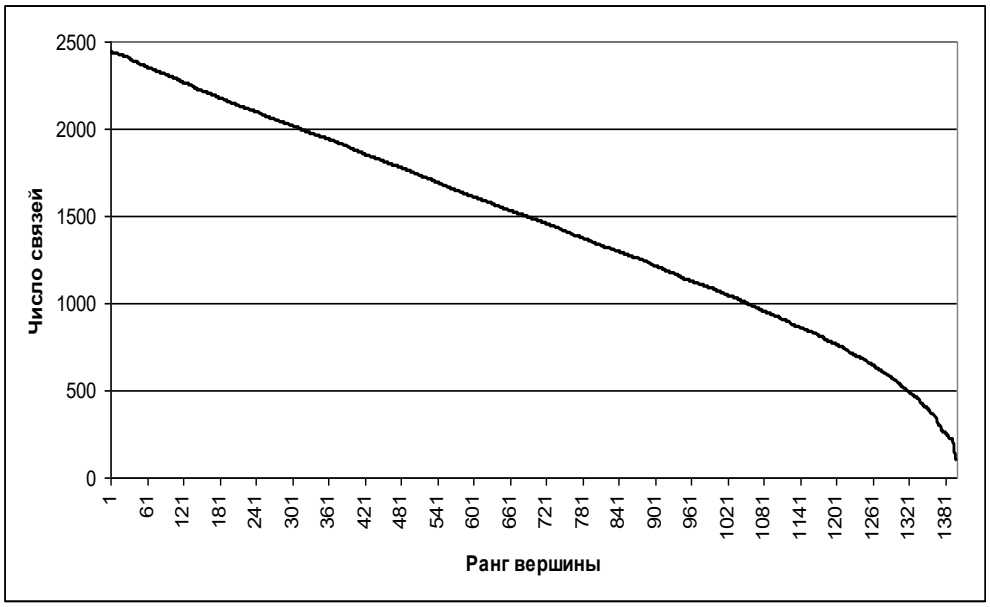

Рис. 1 – Зависимость степени вершины от ранга

Статистические характеристики этого распределения следующие.

1. Наиболее вероятное значение числа связей у города 2439, таких городов в системе 15, соответствующий ранг этих городов равен 2.

2. Медиана распределения составляет 1550 связей, ей отвечает ранг 645.

3. Средняя степень вершины равна 1537, чему отвечает ранг 655.

4. Стандартное отклонение степени вершины равно 732.

5. Стационарная точка распределения степени вершины равна 1518, что близко к медиане и характерно для равномерного распределения. Этой точке отвечает ранг города, равный 670.

6. Минимальное число ребер у вершины равно 97, ранг этого города равен 1400, такой город один, все его связи – с вершинами первых рангов.

7. Максимальное число ребер у вершины равно 2440 (из 2441 городов), т.е. город первого ранга связан со всеми прочими городами в системе. Таких городов первого ранга имеется 7.

Функция распределения вершин по рангу близка к линейной. Это означает, что за исключением лидеров, связанных со всеми, и периферии, имеющей малое количество связей, ранги остальных городов равновероятны. Тем самым структура городских связей имеет естественное деление на три иерархических класса: полносвязное крупное ядро, промежуточную зону и относительно малую периферию. Периферию образуют города с рангами от последнего (1400-го) до 1252-го, куда входят 182 города со связями от 97 до 666. Совокупно эти города составляют 0,075 от общего числа городов системы, так что периферия содержится в последнем дециле распределения городов по степеням вершин, точнее – в третьем квартиле последнего дециля. Обращаясь к рис. 1, видим, что примерно в этой точке линейная зависимость переходит в корневую.

Города первых рангов связаны между собой, т.е. они образуют полный подграф. Количество городов, с которыми нет связи у городов первых рангов,

линейно возрастает с ростом ранга, что позволяет оценить максимальный полный подграф по максимальному количеству связей последнего по рангу входящего в него города. Это количество приближенно оценивается из линейной связи степени вершины с порядковым номером города. При возрастании ранга степень вершин, как было показано выше, уменьшается почти линейно. В этой связи удобно считать идеальную модель рассматриваемой системы дружеских связей в виде линейной зависимости числа связей от ранга (номера) вершины.

Рассмотрим модель равномерно распределенных степеней вершин. Положим, что вершины занумерованы в порядке убывания числа имеющихся связей, линейно зависящих от номера. Первая вершина имеет максимальное число связей $R-1$, а последняя (с номером R) – минимальное, равное $N_{\min}$. Тогда число связей у вершины с номером k равно

$$n(k) = R-1-[c(k-1)],\ \ c = 1-N_{\min}/(R-1). \qquad (14)$$

Здесь квадратные скобки означают целую часть числа, а коэффициент c в нашем случае приблизительно равен $c = 0{,}96$. Формула (14) не полностью определяет структуру графа: требуется еще указать типичные правила соединения вершин. Мы выделяем три области графа в соответствии со структурой матрицы смежности H_k^{γ}.

Первая область отвечает вершинам с большим числом связей; убывание количества связей происходит из-за выключения из орбиты дружбы городов третьей области с малым числом связей. Эта область простирается условно от первого элемента матрицы смежности до границы максимально полного подграфа. Третья область городов с малым количеством дружеских связей характеризуется нулевой диагональной подматрицей, указывающей на то, что города с малым числом связей дружат не между собой, а с городами первых рангов. Между этими двумя областями расположена средняя – вторая зона, где происходит переход от центра к периферии. В идеале этой прослойки нет, и тогда «городской» граф характеризуется двумя типами сообществ. Оценим в этом приближении размерность максимального полного подграфа. Пусть искомая размерность равна k_t. Тогда количество ребер такого упрощенного графа равно

$$K = \frac{k_t(k_t-1)}{2} + (R-k_t)(2k_t-R) + \frac{(R-k_t)(R-k_t+1)}{2}.$$

Пренебрегая для простоты единицей по сравнению количеством вершин R и k_t и вводя величину $x = k_t/R$, получаем отсюда уравнение относительно x:

$$x^2 - 2x + (1+\rho)/2 = 0,$$

где $\rho = 2K/R^2$ есть средняя плотность всего графа. Следовательно, максимальная размерность полного подграфа равна

$$k_t = R\left(1-\sqrt{\frac{1-\rho}{2}}\right). \qquad (15)$$

Для рассматриваемой системы городов получаем $x = 0{,}57$ и $k_t = 1390$. На долю этого подграфа приходится примерно 970 тыс. ребер или 52 % всех связей графа, что довольно близко к реальности (55 % связей графа образуют полный подграф). В идеальной модели других непересекающихся с этим подграфом полных подграфов нет.

В реальной системе существует непустая прослойка в виде второй зоны, где спорадически рассеяны полные подграфы малых размерностей (3, 4, 5, …), образованные вершинами с промежуточным числом связей. Эти промежуточные вершины связаны не только с центром графа, но еще и между собой. Это уменьшает фактическую размерность нулевой подматрицы в последних рангах. Однако, как показали расчеты [12], отличие данной модели от факта в области периферии составляет 1 %, т.е. пренебрежимо мало. Тем самым применяемая модель «городского» графа в виде двух зон – полного подграфа и связанной только с ним периферии – достаточно адекватна.

На рис. 2 приведен граф возрастных связей без учета пола и региона проживания.

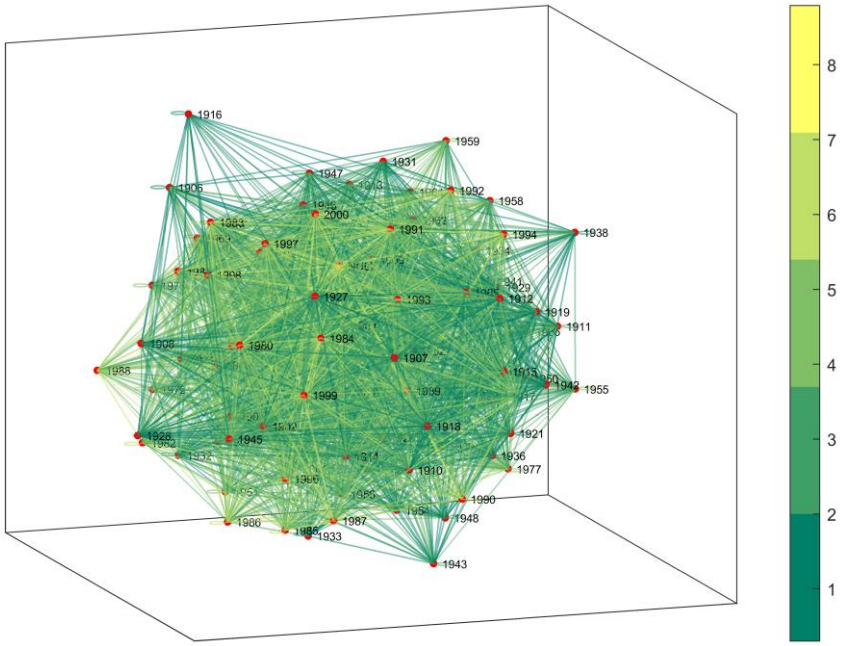

Рис. 2 – Граф межвозрастных связей

В легенде справа показана цветовая индикация количества отдельных пользователей в терминах показателей степеней десятки. Вершины отмечены годом рождения пользователей. Плотность этого графа равна $\rho_A = 0{,}487$.

На рис. 3 показана зависимость степени вершины «возрастного» графа от ранга. Графики на рис. 3 и рис. 1 похожи: в основной части каждый из них с высокой точностью аппроксимируется линейной зависимостью.

Что касается распределения вершин «возрастного» графа по степеням, то, в отличие от «городского», оно не равномерно. Наиболее вероятно число связей

между пятьюдесятью возрастами. Связи в количестве, меньшем 40 и большем 60, в сумме составляют примерно 15 % распределения. Интересно, что даже при наличии 100 млн пользователей ни одна из вершин не связана со всеми возрастными вершинами. Наибольший охват дружбы по возрастам составляет примерно 0,6 возрастного диапазона, тогда как у городских связей, как мы видели, есть примеры полного охвата.

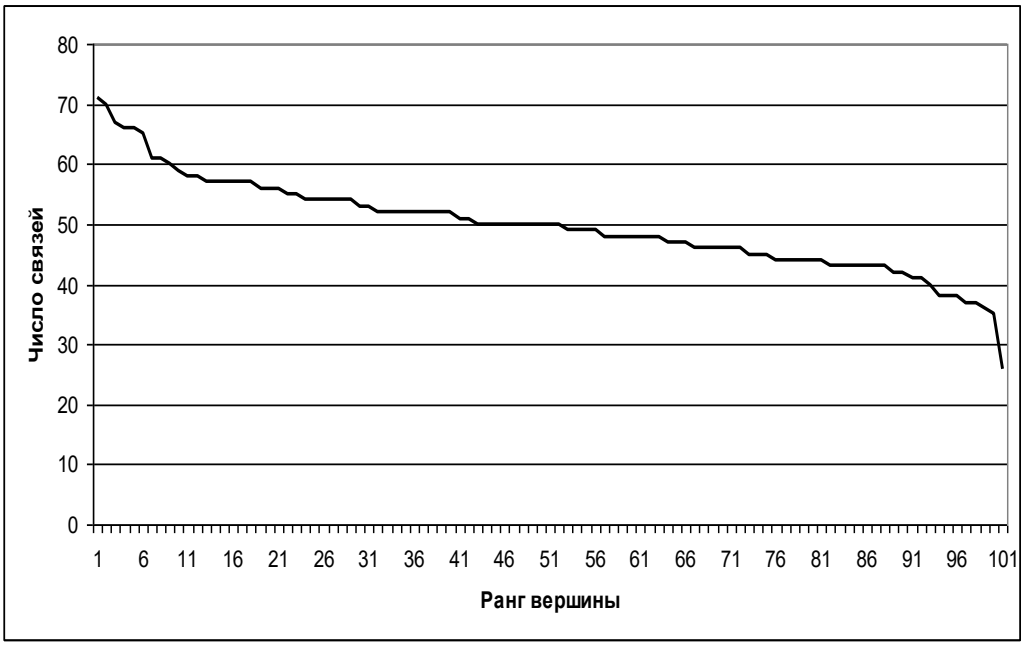

Рис. 3 – Зависимость степени вершины «возрастного» графа от ранга

Функциональная зависимость числа связей от ранга вершины на рис. 3 приближенно разбивается на три линейных участка: первые 10 рангов отвечают уменьшению числа связей на единицу с увеличением ранга на единицу; последним 10 рангам отвечает аналогичное уменьшение на 3 связи на 2 ранга; средние 80 рангов вершин соответствуют уменьшению на 1 связь при уменьшении ранга на 4 единицы.

Матрица смежности F_j^{β} «возрастного» графа имеет существенно иную структуру, чем матрица H_k^{γ} «городского» графа. Если городской граф имеет большое полносвязное ядро, то в возрастном графе максимальная размерность полного подграфа равна 7, таких подграфов довольно много, и они слабо – через одну-две вершины – связаны между собой.

Фактическое число связей по возрасту и городу зависит от рангов городов k и γ. Если зафиксировать ранг k города и просуммировать связи по остальным городам, получим возрастное распределение связей для города k. Для города с большим числом связей (первых рангов в городской системе) матрица смежности $Q_j^{\beta}(k)$, отвечающая распределению связей $q_j^{\beta}(k)$, образует связный граф. Его плотность равна 0,44, что лишь на 0,05 меньше плотности «возрастного» графа, отвечающего всей совокупности регионов. Расстояния между матрицей смежности «возрастного» графа для всей системы и

матрицами $Q_j^{\beta}(k)$ для городов первого, среднего и последнего рангов оказались равными согласно формуле (9) величинам 0,31, 0,88 и 0,98.

Также отметим, что для городов среднего ранга характерно наличие изолированных точек (рис. 4 (а)), отвечающих связям между пользователями одного возраста, а у городов из зоны периферии кроме изолированных точек существуют несвязные подграфы из большего числа вершин (Рис. 4 (б)). Чем выше ранг города, тем выше вероятность того, что граф с матрицей смежности $Q_j^{\beta}(k)$ является связным.

Кроме того, несмотря на достаточно большое количество пользователей даже в городе последнего ранга, только первые 55 городов полносвязного городского ядра имеют связи между всеми возрастами. Города последующих рангов имеют меньшее число связей между меньшим количеством возрастов. Это убывание линейно зависит от ранга.

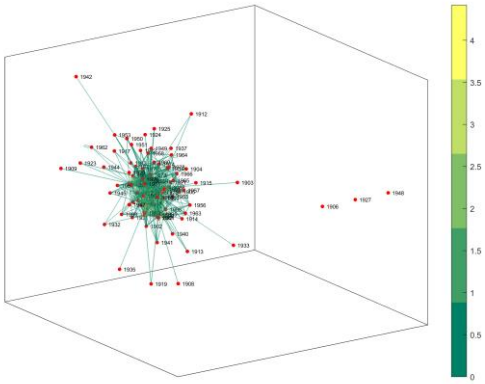

Рис. 4 (а) – Возрастные связи города среднего ранга

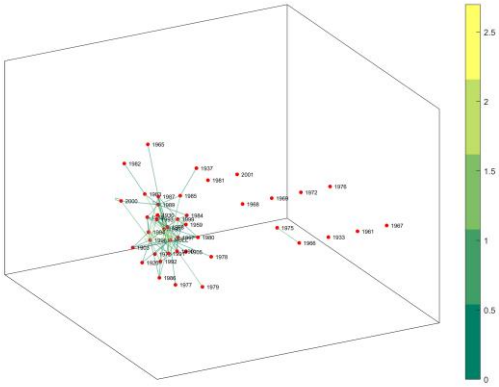

Рис. 4 (б) – Возрастные связи города из периферии

Что касается связей пользователей из городов определенных рангов со всеми остальными пользователями сети в определенном возрастном диапазоне, то здесь вид распределения вероятностей для всех регионов достаточно близок. Для удобства разобъем возрастной промежуток на следующие классы: 1-й ($a1$) от 1902 до 1945 года рождения; следующие 7 классов ($a2$-$a8$) идут с шагом в 5 лет до 1985 г.; затем идут 5 классов ($a9$-$a13$) с шагом в 2 года до 1997 г.; после этого каждый год от 1998 до 2001 образует отдельный класс. Последним 18-м классом является вариант «не указанный год». Выяснилось, что независимо от ранга города наиболее вероятными являются связи с пользователями 10-го возрастного диапазона – это возраст 25-26 лет.

Примерно 55 % всех связей между данным возрастом относится к связям с неуказанным возрастом. В то же время максимум связей пользователей с неуказанным возрастом приходится на возрастной диапазон 10. Тогда можно сделать предположения: основная часть пользователей с неуказанным возрастом относится к диапазону 10; основная часть пользователей с возрастами из первых классов (условно «столетних») в действительности относится к диапазону 10.

В результате получаем следующую модель для матрицы смежности «возрастного» графа $Q_j^{\beta}(k)$ в зависимости от «городского» ранга k. Пренебрегая изолированными вершинами для городов с малым числом связей, считаем, что главным образом структура графа определяется связным подграфом наибольшей размерности. В терминах введенных 18-ти укрупненных классов город с малым числом связей содержит полное возрастное ядро, состоящее из классов $a10$, $a11$, $a12$, $a13$, $a18$. Половина всех связей приходится на ребра, соединяющие $a18$ с $a18$. Уменьшение ранга города, т.е. увеличение числа его связей, приводит к появлению дополнительных полных подграфов, причем ненулевые элементы матрицы смежности начинают появляться сначала в обрамлении выделенного полного подграфа. Связи между возрастами первых классов имеются только для городов первых рангов. Для города среднего ранга характерна полносвязная структура в классах $a4 \div a16, a18$, то есть первая половина городов имеет достаточно плотную структуру возрастных связей, а вторая является разреженной.

5. Заключение

В результате проведенного анализа сети «ВКонтакте» дружеских связей между городами РФ можно сделать следующие выводы.

1. Граф связей плотный, не содержащий узких мест типа «бутылочное горло». Он состоит из полносвязного ядра, состоящего примерно из половины городов, и связанных с ним городов периферии, которые почти не связаны между собой.

2. Распределение вершин по количеству связей почти равномерное, что позволяет аналитически оценить размерность полного подграфа и периферии.

3. Для анализа эволюции структуры графа применено демографическое уравнение. Его можно использовать при мониторинге сетей для оценки на его основе переходных вероятностей и источниковых членов, отвечающих локальному по времени изменению социальной обстановки. Этот подход предполагается развить в последующих работах.

Литература

1. Батура Т.В. Методы анализа компьютерных социальных сетей // Вестник НГУ, 2012. Т.10, вып. 4. С. 13-28.
2. Коломейченко М.И., Чеповский А.Н. Визуализация и анализ графов больших размеров // Бизнес-информатика, 2014. №4. С. 7-16.
3. Гусарова Н.Ф. Анализ социальных сетей. Основные понятия и метрики. – СПб.: ИТМО, 2016. – 67 с.
4. Erdosh P., Renyi A., Sos V.T. On a problem of graph theory // Studia Sci. Mat. Hungar, 1966. V.1. P. 215-235.
5. Barabasi L.-A., Albert R. Emergence of scaling in random networks // Science, 1999. V. 286. P. 509-512.
6. Leskovec J., Chakrabarti D., Kleinberg J., Faloutsos C., Gharamani Z. Kronecker graphs: an approach to modeling networks // J. Machine Learning Research, 2010. V. 11. P. 985-1042.
7. Райгородский А.М. Модели Интернета. – Долгопрудный, Издательский Дом «Интеллект», 2013. – 64 с.
8. Domenico M., Lancichinetti A., Arenas A., Rosvall M. Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems // Phys. Rev., 2015. Vol. X 5. P. 011027.
9. Виртуальное население России // http://webcensus.ru/
10. Чекмышев О.А., Яшунский А.Д. Извлечение и использование данных из электронных социальных сетей // Препринты ИПМ им. М.В. Келдыша. 2014. № 11. 16 с. URL: http://library.keldysh.ru/preprint.asp?id=2014-62
11. Замятина Н.Ю., Яшунский А.Д. Виртуальная география виртуального населения // Мониторинг общественного мнения: Экономические и социальные перемены, 2018. № 1. С. 117-137
12. Кислицын А.А., Орлов Ю.Н. Структура сильно связной компоненты сетевого графа // Препринты ИПМ им. М.В. Келдыша. 2020. № 27. 16 с.