УДК 681.3.06

СЛУЧАЙНЫЕ ГРАФЫ С НЕЛИНЕЙНЫМ ПРАВИЛОМ ПРЕДПОЧТИТЕЛЬНОГО СВЯЗЫВАНИЯ

В.Н. Задорожный

Установлено точное распределение локальной степени связности вершин в случайных графах, выращиваемых по нелинейному правилу предпочтительного связывания. Разработаны методы калибровки генераторов случайных графов.

Ключевые слова: случайные графы, большие сети, моделирование.

ВВЕДЕНИЕ

Классический случайный граф на N вершинах определяется процедурой его построения: каждая пара вершин случайно и независимо с заданной вероятностью p соединяется ребром [1]. Характеристики такого графа выражаются через его параметры N и p. Локальная степень k связности его вершин имеет биномиальное распределение вероятностей со средним $\langle k\rangle = p(N-1)$. При $\langle k\rangle = \text{const}$ и $N\to\infty$ распределение степени вершин графа сходится к пуассоновскому распределению. Классический случайный граф хорошо изучен, однако для моделирования ряда реальных сетей корректно применить его не удается. Это привело к появлению большого числа работ, посвященных исследованию других видов случайных графов, так же (как правило) определяемых процедурами их построения.

Многие изучаемые прикладными науками сети, состоящие из миллионов элементов, развиваются из небольших сетей путем неограниченного наращивания их новыми узлами с постоянным или случайным числом связей. Такие сети моделируются случайными графами, которые выращиваются с помощью простых алгоритмов (генераторов), воспроизводящих способы развития моделируемых сетей. В теории scale-free (безмасштабных) сетей [2—6] наиболее широко известен граф Барабаши—Альберт (БА-граф), предложенный в работе [2], который выращивается из графа-затравки путем циклического добавления к нему новых вершин с постоянным числом $m \geqslant 1$ ребер. В соответствии с правилом предпочтительного связывания («богатые становятся богаче») вероятность p_i связывания нового ребра с i-й вершиной графа пропорциональна ее локальной степени связности k_i:

$$p_i = k_i/\Sigma_j k_j. \qquad (1)$$

Для бесконечного БА-графа, выращиваемого по правилу предпочтительного связывания (1), в работе [5] найдено точное распределение степени связности вершин: вероятность Q_k того, что случайно выбранная вершина имеет степень k, определяется выражением

$$Q_k = \frac{2m(m+1)}{k(k+1)(k+2)}, \quad k = m,\ m+1,\ \ldots \qquad (2)$$

Асимптотически степенные при $k\to\infty$ («масштабно инвариантные») распределения степени связности, обнаруживаемые в больших сетях, обусловливают их специфические свойства и отражаются в названии теории scale-free сетей. Развитый аппарат асимптотического анализа таких сетей позволяет эффективно исследовать их, не привлекая точные решения, подобные (2). Однако для анализа свойств, обусловленных узлами с малыми степенями k и для анализа сетей, имеющих существенно не степенное распределение степени связности узлов, нужны точные решения. В статье выводится точное распределение вероятностей степени вершин для графа, выращиваемого по нелинейному правилу предпочтительного связывания, которое определяет вероятность связывания p_i в виде:

$$p_i = f(k_i)/\Sigma_j f(k_j), \qquad (3)$$

где $f(k) = f_k \geqslant 0$ — вес (ценность выбора) вершины со степенью k — произвольная функция от k. На

основе полученных формул разрабатываются методы калибровки генераторов, позволяющие выращивать графы с заданным распределением степени связности вершин.

1. СТАЦИОНАРНОЕ РАСПРЕДЕЛЕНИЕ СТЕПЕНИ СВЯЗНОСТИ

1.1. Вывод формул распределения

Рассмотрим процесс выращивания графа генератором со стохастическим приращением, представляющим собой добавляемую вершину со случайным числом x ($g \leqslant x \leqslant h$) ребер. Будем считать, что в качестве «затравки» используется конечный граф, для которого выражение $\Sigma_j f(k_j)$ в формуле (3) не равно нулю (он должен содержать вершины со степенями k, для которых $f(k) > 0$) и все $k_j \geqslant g$. На каждом шаге генерации к имеющимся N вершинам графа добавляется новая вершина и связывается x ребрами с вершинами графа, выбираемыми случайно в соответствии с правилом (3). Тогда с ростом числа N вершин выращиваемого графа при их минимальной степени g средняя степень сходится к величине $\langle k\rangle = 2m$, где m $\langle x\rangle$ — математическое ожидание x.

Обозначим через r_k ($k \geqslant g$) вероятность того, что приращение содержит k ребер. Предполагая, что при $N \to \infty$ существует стационарный режим развития графа, выпишем уравнения баланса для финальных вероятностей Q_k степеней вершин. Для этого определим слой A_k как множество вершин графа, имеющих степень k. Вероятность q_k слоя A_k определим как вероятность того, что случайно (равновероятно) выбранная вершина принадлежит слою A_k: $q_k = |A_k|/N$, где $|A_k|$ — число вершин со степенью k. Вероятность Q_k определим как финальное значение вероятности q_k при $N \to \infty$.

Финальное значение P_k вероятности $\sum\limits_{i \in A_k} p_i$ того, что для связывания с ребром новой вершины будет выбрана вершина слоя A_k, $k \geqslant g$, найдем, исходя из правила (3):

$$\sum_{i \in A_k} p_i = \frac{\sum\limits_{i \in A_k} f(k_i)}{\sum\limits_{j=1}^{N} f(k_j)} = \frac{f(k)|A_k|}{\sum\limits_{l \geq g} f(l)|A_l|} = \frac{f_k|A_k|/N}{\sum\limits_{l \geq g} f_l|A_l|/N} \sim$$
$$\sim \frac{Q_k f_k}{\sum\limits_{l \geq g} Q_l f_l} = \frac{Q_k f_k}{\langle f\rangle} = P_k, \qquad (4)$$

где $\langle f\rangle = \sum\limits_{l \geq g} Q_l f_l$ — средний вес вершины графа.

Финальные вероятности Q_k найдем из уравнений баланса, которые выпишем следующим образом. За один шаг генерации в граф добавляется одна вершина с x ребрами. Число $|A_g|$ вершин слоя A_g увеличивается в среднем на r_g (в этот слой с вероятностью r_g попадает новая вершина) и одновременно уменьшается (так как из этого слоя уходят вершины, когда к ним присоединяется новое ребро) в среднем на $mP_g = mQ_g f_g/\langle f\rangle$ вершин. Первое уравнение баланса получаем, приравнивая выражения для Q_g до и после шага генерации в стационарном режиме развития графа (при $N \to \infty$):

$$\frac{|A_g|}{N} = \frac{|A_g| + r_g - mQ_g f_g/\langle f\rangle}{N+1}.$$

Отсюда $|A_g|N + |A_g| = N|A_g| + Nr_g - NmQ_g f_g/\langle f\rangle$ или $|A_g|/N = r_g - mQ_g f_g/\langle f\rangle$. Учитывая, что $|A_g|/N \sim Q_g$, получаем $Q_g = r_g - mQ_g f_g/\langle f\rangle$ и находим:

$$Q_g = \frac{r_g\langle f\rangle}{\langle f\rangle + mf_g}, \qquad (5)$$

где f_g — вес вершины с наименьшей степенью связности.

Аналогично выписываем уравнения баланса при $k > g$. Число $|A_k|$ вершин слоя A_k возрастает за шаг генерации в среднем на $r_k + mP_{k-1} = r_k + mQ_{k-1} \times$ $\times f_{k-1}/\langle f\rangle$ (за счет попадания с вероятностью r_k новой вершины в слой A_k и прихода вершин из слоя A_{k-1}), и уменьшается в среднем на $mP_k = mQ_k f_k/\langle f\rangle$ (за счет ухода вершин из слоя A_k в слой A_{k+1}). Получаемое уравнение

$$\frac{|A_k|}{N} = \frac{|A_k| + r_k + \dfrac{mQ_{k-1}f_{k-1}}{\langle f\rangle} - \dfrac{mQ_k f_k}{\langle f\rangle}}{N+1}$$

решаем аналогично первому уравнению баланса и находим:

$$Q_k = \frac{r_k\langle f\rangle + mf_{k-1}Q_{k-1}}{\langle f\rangle + mf_k}, \quad k \geqslant g+1. \qquad (6)$$

При $m = g$ (при $r_g = 1$) стохастическое приращение графа вырождается в постоянное, т. е. в вершину с постоянным числом $m \geqslant 1$ ребер. Найденное для него в работе [7] распределение степени вершин является соответствующим частным случаем распределения (5), (6):

$$Q_g = \frac{\langle f\rangle}{\langle f\rangle + mf_g}, \qquad (7)$$

$$Q_k = \frac{mf_{k-1}}{\langle f\rangle + mf_k} Q_{k-1}, \quad k \geqslant g+1. \qquad (8)$$

Если $f_k = k$, то $\langle f \rangle = \langle k \rangle = 2m$, $f_g = g = m$, и $Q_g = \frac{2}{2+m}$, $Q_k = \frac{k-1}{2+k} Q_{k-1}$, $(k \geqslant g+1)$. Отсюда индукцией по k легко выводится известная формула (2).

Заметим, что весовую функцию $f(k) = f_k$, как это видно из формулы (3), можно определять с точностью до мультипликативной константы, и, следовательно, веса cf_k $(c > 0)$ индуцируют то же распределение Q_k, $k \geqslant g$ что и веса f_k.

1.2. Численный расчет распределения степеней

Непосредственное использование рекурсивных формул (5), (6) для численного расчета вероятностей Q_k затрудняется тем, что при известных m, r_k и f_k средний вес $\langle f \rangle$ неизвестен. Его можно определять из уравнения, получаемого в результате подстановки (5), (6) в равенство $\langle k \rangle = \sum_{k \geq g} kQ_k = 2m$. Такой расчет $\langle f \rangle$ легко реализуется в среде электронных таблиц следующей процедурой.

1. Формируем столбцы значений k, f_k, r_k и ячейку m.

2. В отдельную ячейку таблицы в качестве начального приближения $\langle f \rangle$ вводим произвольное значение $a > 0$.

3. Со ссылкой на ячейку a как на параметр $\langle f \rangle$ по рекурсивным формулам (5), (6) формируем столбец вероятностей Q_k (его начальное приближение).

4. Вводим в отдельные ячейки:

— формулу, вычисляющую $\langle k \rangle$ как сумму произведений столбцов k и Q_k,

— формулу, вычисляющую $\langle f \rangle$ как сумму произведений столбцов f_k и Q_k.

5. Используя сервис «Подбор параметра», находим искомое a, доставляющее равенство ячейки $\langle k \rangle$ значению $2m$. При таком a автоматически выполнится и равенство ячейки $\langle f \rangle$ значению a. В результате одновременно получаем искомые значения $\langle f \rangle$ и Q_k.

При расчете вероятностей (7), (8) эта процедура упрощается: столбец r_k в этом случае не формируется, а столбец Q_k рассчитывается, соответственно, по формулам (7) и (8).

2. СВОЙСТВА РАСПРЕДЕЛЕНИЙ, РЕАЛИЗУЕМЫХ ПРИ ПОСТОЯННОМ ПРИРАЩЕНИИ

2.1. Общая формула вероятностей

При анализе свойств распределения вероятностей (7), (8) весовую функцию $f(k) = f_k$ удобно рассматривать как произвольный упорядоченный набор чисел $\mathbf{f}_g = f_g, f_{g+1}, ..., f_k, ...$, в котором $f_g > 0$ и $f_k \geqslant 0$ при $k \geqslant g$.

Как можно видеть из формулы (8), первым в наборе $\mathbf{f}_g$ нулевым значением $f_{M+1} = 0$, $M \geqslant g$, предопределяется равенство $Q_k = 0$ для всех $k > M+1$. Поэтому веса $\mathbf{f}_g$ будем задавать либо конечным набором $\mathbf{f}_g^M = f_g, f_{g+1}, ..., f_M, 0$, все числа которого, кроме последнего $f_{M+1} = 0$, положительны, либо бесконечным набором $\mathbf{f}_g^\infty$ положительных чисел. В сокращенной записи $\mathbf{f}_g^M = f_g, f_{g+1}, ..., f_M$, содержащей лишь положительные числа, вес f_{M+1} равен нулю по умолчанию.

При некотором $a \geqslant 0$ конечный набор весов $\mathbf{f}_g^M$ индуцирует по формулам (7), (8) распределение вероятностей $\mathbf{Q}_g^{M+1} = Q_g, Q_{g+1}, ..., Q_{M+1}$, в котором

$$Q_g = \frac{a}{a + mf_g},$$

$$Q_{g+1} = \frac{a}{a + mf_g} \frac{mf_g}{a + mf_{g+1}}, ..., Q_{M+1} =$$

$$= \frac{a}{a + mf_g} \frac{mf_g}{a + mf_{g+1}} ... \frac{mf_M}{a + mf_{M+1}},$$

или, в более компактном виде:

$$Q_k = \frac{a}{a + mf_g} \prod_{i=g+1}^{k} \frac{mf_{i-1}}{a + mf_i}, \quad g \leqslant k \leqslant M+1. \quad (9)$$

Произведение $\prod_{i=g+1}^{g}$ в котором верхний предел меньше нижнего, здесь равно единице по определению. Так как $f_{M+1} = 0$, то для Q_{M+1} имеем также выражение

$$Q_{M+1} = \frac{a}{a + mf_g} \cdot \frac{mf_g}{a + mf_{g+1}} ... \frac{mf_{M-1}}{a + mf_M} \cdot \frac{mf_M}{a + mf_{M+1}} =$$

$$= \prod_{i=g}^{M} \frac{mf_i}{a + mf_i}, \quad (10)$$

в котором при $a = 0$ не возникает неопределенности типа 0/0.

Для случая $a = 0$ из формулы (9) с учетом выражения (10) при любых ненулевых весах $\mathbf{f}_g^M$ получаем:

$$\mathbf{Q}_g^{M+1} = Q_g, ..., Q_M, \quad Q_{M+1} = 0, ..., 0{,}1. \quad (11)$$

Здесь $\langle f\rangle = f_g Q_g + \ldots + f_{M+1} Q_{M+1} = 0 = a$, и случайная величина k вырождается в константу $k = M + 1$. Условие реализуемости случая $a = 0$ и распределения (11) уточняется далее в п. 2.4.

Полагая $M \to \infty$, из (9), (10) находим, что в общем случае набор $\mathbf{f}_g^\infty$ индуцирует распределение $\mathbf{Q}_g^\infty = Q_g, Q_{g+1}, \ldots, Q_k, \ldots$, в котором вероятность Q_∞, определяемая как предел произведения (10) при $M \to \infty$, может быть положительной.

2.2. Индуцируемые распределения

Покажем, что для ряда $\mathbf{Q}_g$ (9) условие нормировки $\sum\limits_{k \geq g} Q_k = 1$ на самом деле выполняется при любых $a \geqslant 0$ и, таким образом, с учетом свойства $\mathbf{Q}_g \geqslant 0$, формула (9) определяет параметрический класс распределений с параметром a.

При $M < \infty$ условие нормировки проверяется непосредственно:

$$\sum_{k \geq g} Q_k = \frac{a}{a+mf_g} + \frac{a}{a+mf_g}\frac{mf_g}{a+mf_{g+1}} + \ldots$$

$$\ldots + \frac{a}{a+mf_g}\frac{mf_g}{a+mf_{g+1}} \ldots \frac{mf_{M-1}}{a+mf_M}\frac{mf_M}{a+mf_{M+1}} =$$

$$= \frac{a}{a+mf_g}\left(1 + \frac{mf_g}{a+mf_{g+1}}\left(1 + \ldots + \frac{mf_{M-2}}{a+mf_{M-1}} \times \right.\right.$$

$$\left.\left. \times \left(1 + \frac{mf_{M-1}}{a+mf_M}\left(1 + \frac{mf_M}{a}\right)\right)\ldots\right)\right) =$$

$$= \frac{a}{a+mf_g}\left(1 + \frac{mf_g}{a+mf_{g+1}}\left(1 + \ldots + \frac{mf_{M-2}}{a+mf_{M-1}} \times \right.\right.$$

$$\left.\left. \times \left(1 + \frac{mf_{M-1}}{a+mf_M}\left(\frac{a+mf_M}{a}\right)\right)\ldots\right)\right) = \ldots$$

$$\ldots = \frac{a}{a+mf_g}\left(\frac{a+mf_g}{a}\right) = 1.$$

В приведенной выкладке учитывается, что $f_{M+1} = 0$.

Выполнение условия $\sum\limits_{k \geq g} Q_k = 1$ для $\mathbf{Q}_g^\infty = Q_g, Q_{g+1}, \ldots$ вытекает из определения Q_∞ как предела вероятности Q_{M+1} при $M \to \infty$.

2.3. Точное определение среднего веса $a = \langle f\rangle$

Финальное распределение вероятностей для степени вершин графа, выращиваемого генератором с постоянным приращением, принадлежит классу индуцируемых распределений (9) и определяется значением a, удовлетворяющим условию $\langle k\rangle = \sum\limits_{k=g}^{M+1} kQ_k = 2m$, или, с учетом (9), (10), условию:

$$\sum_{k=g}^{M+1} \frac{ak}{a+mf_g} \prod_{i=g+1}^{k} \frac{mf_{i-1}}{a+mf_i} = 2m \Leftrightarrow$$

$$\Leftrightarrow \sum_{k=g}^{M} \frac{ak}{a+mf_g} \prod_{i=g+1}^{k} \frac{mf_{i-1}}{a+mf_i} +$$

$$+ (M+1)\prod_{i=g}^{M} \frac{mf_i}{a+mf_i} = 2m. \qquad (12)$$

Условие (12) в любой из его двух эквивалентных форм можно использовать как уравнение для среднего веса $a = \langle f\rangle$ в распределении (7), (8). Левая часть уравнения (12) в обеих его формах есть функция аргумента a, в промежутке $0 \leqslant a < \infty$ монотонно убывающая от $M + 1$ до $g = m$. Это гарантирует существование и единственность решения при $M + 1 > m$ и позволяет эффективно находить его численным методом, описанным в п. 1.2.

2.4. Условие существования стационарного режима. Псевдорешетки

Назовем вакансией еще неиспользованную возможность связывания вершины с новым ребром. В случае конечного набора весов $\mathbf{f}_g^M$ число вакансий у любой вершины ограничено и составляет $M + 1 - k$, где k — степень связности этой вершины. Для выращивания бесконечного графа необходимо, чтобы число вакансий, добавляемых в граф новой вершиной, было не меньше числа вакансий, используемых ее ребрами, т. е. должно выполняться условие

$$M + 1 - m \geqslant m \text{ или } M + 1 \geqslant 2m. \qquad (13)$$

При $M + 1 = 2m$, на границе условия (13), уравнение (12) при любых $f_g, \ldots, f_M > 0$ имеет решение $a = 0$, определяющее вероятности Q_k (11). Так как при $M + 1 = 2m$ число вакансий в растущем графе не изменяется, то вершины бесконечного графа с вероятностью 1 имеют степень $M + 1$. Такой граф можно назвать псевдорешеткой, так как постоянством степени вершин он подобен решетке. Варьируя m и вычисляя $M = 2m - 1$, можно получать разные виды псевдорешеток. При $m = 1$, $M = 1$ генерируется бесконечная цепь со степенями вершин $M + 1 = 2$, при $m = 2$, $M = 3$ степени почти всех вершин равны 4 и т. д.

3. КАЛИБРОВКА ГЕНЕРАТОРОВ С ПОСТОЯННЫМ ПРИРАЩЕНИЕМ

3.1. Задача калибровки

Для заданного распределения $\tilde{\mathbf{Q}}_g^{M+1}$ требуется найти веса $\mathbf{f}_g^M > 0$, при которых степень вершин имеет распределение $\mathbf{Q}_g^{M+1} = \tilde{\mathbf{Q}}_g^{M+1}$.

3.2. Реализация экспоненциального распределения

Дискретное экспоненциальное распределение является геометрическим распределением вероятностей. Из выражений (7), (8) легко видеть, что оно реализуется при использовании равных весов, например при $\mathbf{f}_g^\infty = 1, 1, \ldots$

3.3. Общий метод калибровки: обращение рекурсии

Найти набор весов $\mathbf{f}_g^M$, индуцирующих нужное распределение $\tilde{\mathbf{Q}}_g^{M+1} > 0$, $\sum\limits_{k=g}^{M+1} \tilde{Q}_k = 1$, можно путем обращения зависимости, заданной рекурсивными соотношениями (7), (8). Выражая в них веса f_k через заданные вероятности $Q_k = \tilde{Q}_k$, $k \geqslant g$, получаем:

$$f_g = \frac{a}{m\tilde{Q}_g} - \frac{a}{m}, \quad f_k = \frac{\tilde{Q}_{k-1}}{\tilde{Q}_k} f_{k-1} - \frac{a}{m},$$
$$(k = g+1, \ldots, M), \quad f_{M+1} = 0.$$

Найденные веса имеют общую мультипликативную константу a/m, заменяя которую единицей (и определяя этим равенство $a = m$), находим решение задачи калибровки в виде

$$f_g = \frac{1}{\tilde{Q}_g} - 1, \quad f_k = \frac{\tilde{Q}_{k-1}}{\tilde{Q}_k} f_{k-1} - 1,$$
$$(k = g+1, \ldots, M), \quad f_{M+1} = 0, \qquad (14)$$

позволяющем вычислять все f_k непосредственно по заданным $\tilde{Q}_k$. Равенства (14) вместе с условием стационарности (13) и необходимым условием $\langle k \rangle = 2m$ реализуемости распределения $\tilde{\mathbf{Q}}_g^{M+1}$ составляют достаточное условие реализуемости $\tilde{\mathbf{Q}}_g^{M+1}$. Докажем это.

Действительно, пусть при $M + 1 \geqslant 2m$, $\langle k \rangle = 2m$ правило предпочтения использует веса (14). Тогда стационарное распределение $\mathbf{Q}_m^{M+1}$ существует и определяется формулами (7), (8). При этом средний вес $a = \langle f \rangle = m$ известен, так как значение $a = m$ выбрано при выводе формул (14). Убедимся, что это значение $a = m$ является корнем уравнения (12):

$$\sum_{k=g}^{M+1} \frac{ak}{a+mf_g} \prod_{i=g+1}^{k} \frac{mf_{i-1}}{a+mf_i} =$$
$$= \sum_{k=g}^{M+1} \frac{mk}{m+mf_g} \prod_{i=g+1}^{k} \frac{mf_{i-1}}{m+mf_i} =$$
$$= \sum_{k=g}^{M+1} \frac{k}{1+f_g} \prod_{i=g+1}^{k} \frac{f_{i-1}}{1+f_i} =$$
$$= \sum_{k=g}^{M+1} \frac{k}{1+1/\tilde{Q}_g - 1} \prod_{i=g+1}^{k} \frac{f_{i-1}}{1+(\tilde{Q}_{i-1}/\tilde{Q}_i)f_{i-1} - 1} =$$
$$= \sum_{k=g}^{M+1} \tilde{Q}_m k \prod_{i=g+1}^{k} \frac{\tilde{Q}_i}{\tilde{Q}_{i-1}} = \sum_{k=g}^{M+1} k\tilde{Q}_k = 2m$$

(последнее из равенств совпадает с посылкой $\langle k \rangle = 2m$). Подставляя теперь в формулы (7), (8) известное среднее $\langle f \rangle = a = m$ и веса (14), получаем $Q_k = \tilde{Q}_k$ для всех $k \geqslant g$.

3.4. Реализация равномерного и треугольного распределений

Найдем набор весов $\mathbf{f}_g^M$ для выращивания графа с равномерным распределением $\tilde{\mathbf{Q}}_g^{M+1}$. Поскольку здесь $\tilde{Q}_g = \tilde{Q}_{g+1} = \ldots = \tilde{Q}_{M+1} = 1/n$, где $n = M - g + 2$, то, используя формулы (14), находим: $f_g = 1/\tilde{Q}_g - 1 = (n-1)$, $f_{g+1} = f_g - 1 = (n-2)$, $\ldots$, $f_k = n - (k - g + 1)$, $\ldots$, $f_M = n - (M - g + 1) = 1$, $f_{M+1} = 0$. Таким образом, получен набор весов $\mathbf{f}_g^M = n-1, n-2, \ldots, 2, 1$. Для выполнения условия реализуемости $\langle k \rangle = 2m$, или $(g + M + 1)/2 = 2g$, должно выполняться соотношение $M = 3g - 1$. Так, например, при $g = 3$, $M = 3g - 1 = 8$ и $\mathbf{f}_3^8 = 6, 5, 4, 3, 2, 1$ генерируется граф с распределением степени связности $\mathbf{Q}_g^{M+1} = \mathbf{Q}_3^9 = 1/7, \ldots, 1/7$.

Аналогично находим веса $\mathbf{f}_g^M$, индуцирующие при $M = 3g - 1$ треугольное распределение $\mathbf{Q}_g^{M+1} = Q_g, 2Q_g, \ldots, mQ_g, (m+1)Q_g, mQ_g, \ldots, 2Q_g, Q_g$ (где $Q_g = 1/(g+1)^2$):

$$f_k = \begin{cases} (g+1)^2/(k-g+1) - (k-g)/2 - 1, & g \leq k \leq 2g, \\ (3g-k)/2, & 2g \leq k \leq 3g. \end{cases}$$

Так, генератор с весами $\mathbf{f}_3^8 = 15, 13/2, 10/3, 3/2, 1, 1/2$ реализует граф с распределением степени вершин $\mathbf{Q}_3^9 = 1/16, 2/16, 3/16, 4/16, 3/16, 2/16, 1/16$.

3.5. Задача реализации усеченного степенного распределения

Усеченные степенные распределения степени связности, обнаруживаемые при исследовании больших сетей [6], реализуются генератором с постоянным приращением приближенно.

Рассмотрим в качестве примера реализацию распределения

$$\tilde{Q}_k = ck^{\alpha} = 1{,}677k^{-2}, \quad k = 2, ..., 20. \qquad (15)$$

Множитель $c = 1{,}677$ обеспечивает равенство суммы вероятностей единице. Здесь имеем $g = m = 2$, $M + 1 = 20$ и среднее $\langle k\rangle = 4{,}3574$, близкое к $2m = 4$. Это позволяет реализовать распределение (15) приближенно, задавая веса по формулам (14). Расчет получаемого при этих весах распределения $\mathbf{Q}_2^{20}$ (см. п. 1.2) показывает, что его среднее составляет $\langle k\rangle = 2m = 4$, и что реализуемые вероятности Q_k (рис. 1, см. светлые маркеры) отличаются от заданных формулой (15) значений $\tilde{Q}_k$ (см. рис. 1, черные маркеры). В конце диапазона $2 \leqslant k \leqslant 20$ погрешность реализации заданных вероятностей $\tilde{Q}_k$ достигает 88 %.

В работах по теории scale-free графов используется другой способ: используются веса $f_k = k$ при $g \leqslant k \leqslant M$ и $f_{M+1} = 0$. Получаемое при $g \leqslant k \leqslant M$ асимптотически-степенное распределение Q_k (см. рис. 1, сплошная линия) хорошо согласуется с заданным степенным распределением. Однако значение Q_{M+1} на порядки превосходит заданное. Это происходит потому, что интенсивный приход вершин в слой A_{M+1} при генерации графа не компенсируется их уходом из этого слоя (так как с ними ребра новых вершин не связываются).

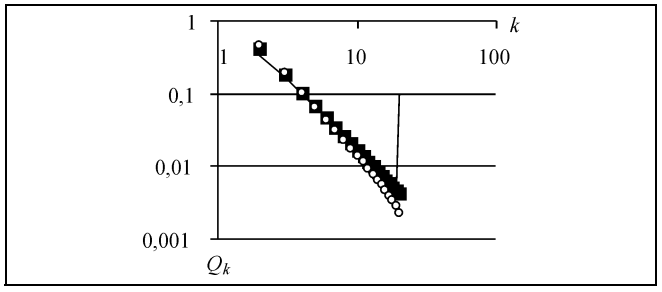

Рис. 1. Графики реализаций распределения (15) (шкалы логарифмические)

Точно реализовать усеченные степенные и любые другие распределения $\tilde{\mathbf{Q}}_g$ со средним $\langle k\rangle \geqslant 2g$ позволяют генераторы со стохастическим приращением.

4. КАЛИБРОВКА ГЕНЕРАТОРОВ СО СТОХАСТИЧЕСКИМ ПРИРАЩЕНИЕМ

4.1. Задача калибровки

Для заданного распределения $\tilde{\mathbf{Q}}_g^{M+1}$ со средним $\langle k\rangle \geqslant 2g$ требуется найти вероятности r_g, ..., r_h, ($r_g + ... + r_h = 1$, $h \leqslant M$) и веса $\mathbf{f}_g^M \geqslant 0$, при которых степень вершин выращиваемого графа будет иметь распределение $\mathbf{Q}_g^{M+1} = \tilde{\mathbf{Q}}_g^{M+1}$.

4.2. Обращение рекурсии

По аналогии с выводом из выражений (7) и (8) формул (14) из выражений (5) и (6) выводятся формулы для калибровки генератора со стохастическим приращением:

$$f_g = \frac{r_g}{\tilde{Q}_g} - 1, \quad f_k = \frac{\tilde{Q}_{k-1}}{\tilde{Q}_k} f_{k-1} + \frac{r_k}{\tilde{Q}_k} - 1,$$
$$(k = g + 1, ..., M), \quad f_{M+1} = 0, \quad \langle f\rangle = m, \qquad (16)$$

где $m = \langle x\rangle = \Sigma k r_k$. При $r_g = 1$ формулы (16) сводятся к формулам (14) и реализуется наименьшее для данного g среднее $\langle k\rangle = 2g$. Вероятности r_g,..., r_h при заданном $\tilde{\mathbf{Q}}_g^{M+1}$ следует задавать так, чтобы выполнялось условие $\langle k\rangle = 2m$. Если генератор использует вычисленные по заданному $\tilde{\mathbf{Q}}_g^{M+1}$ неотрицательные веса (16), выполняется условие стационарности и $\langle k\rangle = 2m \geqslant 2g$, то $a = \langle f\rangle = m$ и генератор реализует распределение $\mathbf{Q}_g^{M+1} = \tilde{\mathbf{Q}}_g^{M+1}$. Справедливость этого утверждения доказывается подстановкой формул (16) в формулы (5) и (6).

Примечание 1. Значения r_g, ..., r_h, выбираемые для выполнения условия $\langle k\rangle = 2m$, должны быть такими, чтобы веса (16) были неотрицательными. Из формул (16) вытекает, что для этого необходимо и достаточно, чтобы при всех $k \leqslant h$ выполнялось условие $\sum_{i=g}^{k} r_i \geqslant \sum_{i=g}^{k} \tilde{Q}_i$. Это позволяет обеспечивать равенство $\langle k\rangle = 2m$ при любом $\langle k\rangle \geqslant 2g$.

Примечание 2. Стационарный режим существует, если $M + 1 \geqslant 2m$. Вероятность p_s его реализации, однако, может быть меньше единицы. Так, при $r_1 = ... = r_4 = 1/4$ и любом $\mathbf{f}_1^4 > 0$ в стационарном

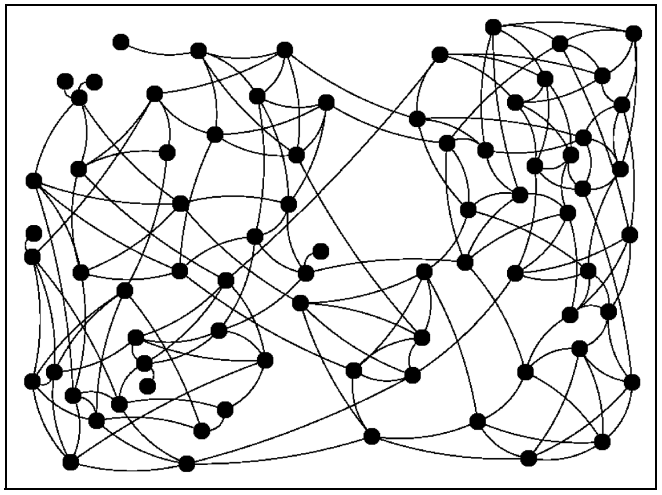

Рис. 2. Начальный фрагмент псевдорешетки, выращиваемой из 10-вершинной цепочки при $\mathbf{f}_1^4$ =1, 10, 100, 1000

режиме (который существует, поскольку $m = 2{,}5$, $M + 1 = 5 = 2m$) псевдорешетка со степенями вершин $k = 5$ (рис. 2) выращивается с вероятностью $p_s = 0$. Действительно, при любом конечном числе $V_0 > 0$ вакансий в затравке их число $V(t)$ в ходе генерации псевдорешетки определяется процессом случайного блуждания на прямой, стартующим из точки V_0. И за конечное число t шагов генерации $V(t)$ возвращается к нулю. В подобных случаях для достоверной реализации стационарного режима те приращения, которые не могут быть реализованы на данном шаге, можно помещать в очередь для их последующей реализации при первой возможности. Тогда при $t \to \infty$ если $M + 1 \geqslant 2m$, то средняя длина $L(t)$ очереди приращений сходится к нулю, а если $M + 1 = 2m$, то $L(t) \to \infty$, но $L(t)/N \sim L(t)/t \to 0$.

Примечание 3. Если $r_k > 0$ при $k = g + 1, \ldots, M$, то требование $f_{k-1} > 0$ можно заменить требованием $f_{k-1} \geqslant 0$. Если при некотором k имеем $r_k > 0$ и $f_k = 0$, то приращения с числом ребер $x = k$, связывая k вакансий в графе, не добавляют новых вакансий, так как с вершинами этих приращений новые ребра связываться не могут. С учетом этого условие стационарности принимает вид $\sum_{k=g}^{h} (M + 1 - k) r_k \omega_k \geqslant m$, где $\omega_k = 0$ при $f_k = 0$ и $\omega_k = 1$ при $f_k > 0$.

Примечание 4. При $f_{k-1} = 0$ слой A_k может формироваться только путем задания $r_k > 0$, т. е. прямым «посевом» вершин со степенью k. При заданном наборе $\mathbf{f}_g^M$ посевом можно формировать слои A_k с номерами k вне отрезка $[g, M + 1]$, в том числе и слой A_0.

4.3. Точная реализация усеченного степенного распределения

Для точной реализации распределения (15) достаточно ввести две положительные вероятности $r_2 = 0{,}8213$ и $r_3 = 0{,}1787$, формируя $m = 2r_2 + 3r_3 = 2{,}1787$, которое удовлетворяет условию $\langle k \rangle = 2m$ реализуемости распределения (15). После этого по формуле (16) рассчитываются веса $\mathbf{f}_2^{19} = f_2, \ldots, f_{19}, 0$ (рис. 3). В компьютерном эксперименте генератор с этими f_k и r_1, r_2 выращивает граф с точным распределением степени связности (15).

4.4. Калибровка генераторов по эмпирическим данным

Публикуемые в Интернете данные о структуре больших сетей позволяют рассчитывать и использовать для калибровки их моделей эмпирические оценки $\hat{\mathbf{Q}}_g$ распределения степени связности.

Оценки $\hat{\mathbf{Q}}_g$, как правило, неоднородны. Обычно начало ряда $\hat{\mathbf{Q}}_g$ представлено оценками $\hat{Q}_k$, которые имеют хорошую точность и соответствуют установившимся стационарным вероятностям Q_k, зависящим от k специфическим нелинейным образом. Середина ряда $\hat{\mathbf{Q}}_g$ требует сглаживания оценок их теоретической зависимостью от k. Оценки в хвостовой части ряда имеют высокую относительную погрешность, содержат длинные цепочки нулей и соответствуют не стационарным вероятностям q_k нарождающихся слоев.

Пошаговое описание методики калибровки генераторов по эмпирическим данным приведем на примере данных [8] о сети автономных систем Интернет (22 963 узла).

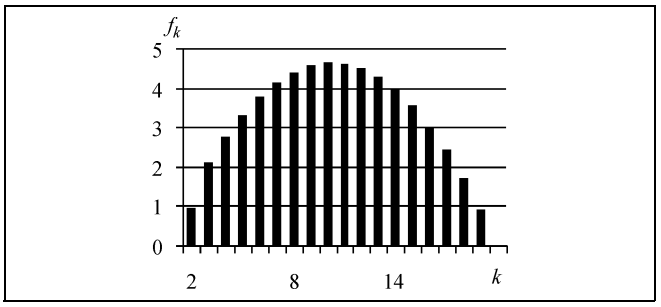

Рис. 3. График весовой функции, индуцирующей распределение (15)

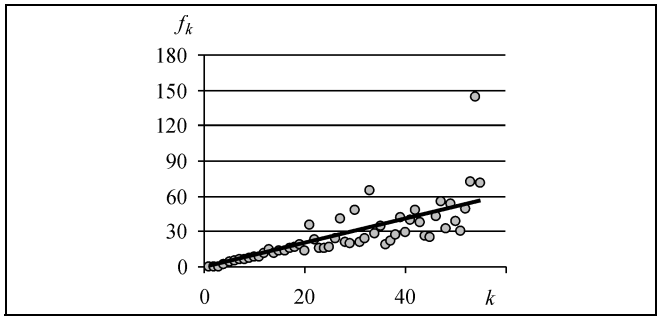

Рис. 4. Определение весовой функции для сети автономных систем Интернет

Шаг 1. После удаления из графа исходной сети [8] двух «подозрительных» вершин с петлями и расчета частот n_k (числа узлов со степенью k) вычисляем оценки $\hat{Q}_k = n_k/N$ (где $N = 22961$, $1 \leqslant k \leqslant 1713$). Например, первые оценки $\hat{Q}_1$, ..., $\hat{Q}_{10}$ равны, соответственно, 0,3414; 0,4225; 0,0966; 0,0395; 0,0205; 0,0126; 0,0087; 0,0075; 0,0057; 0,0049.

Шаг 2. Вычисляем оценки $\hat{v}_k = \sqrt{\hat{Q}_k(1 - \hat{Q}_k)/n_k}\,/\,\hat{Q}_k$ коэффициентов вариации оценок $\hat{Q}_k$. Учитывая, что при $k \leqslant 4$ вариации $\hat{v}_k$ невелики ($\hat{v}_k < 0{,}2$), принимаем решение использовать оценки $\hat{Q}_1$, ..., $\hat{Q}_4$ как «истинные» вероятности Q_1, ..., Q_4.

Шаг 3. Определяем по данным [8] среднее число m ребер в приращении и среднюю степень $\langle k \rangle$ связности: $m = R/N = 2{,}014$ (где R — число ребер в сети), $\langle k \rangle = 2R/N = 4{,}027$.

Шаг 4. Задаем для стохастического приращения минимальное число положительных вероятностей r_k, обеспечивающих требуемое m при положительных f_k (16). В данном примере, задавая $r_1 = 0{,}342$, $r_2 = 0{,}432$, $r_3 = 0{,}096$, $r_4 = 0{,}13$, имеем $m = r_1 + 2r_2 + 3r_3 + 4r_4 = 2{,}014$ и f_1, ..., $f_4 = 0{,}0017$, $0{,}0245$, $0{,}0999$, $2{,}5303$ соответственно.

Шаг 5. По набору оценок $\hat{Q}_1$, ..., $\hat{Q}_{55}$, ограниченному первым нулем $\hat{Q}_{56} = 0$, используя формулы (16), находим веса f_1, ..., f_{55}, изображаем их на графике (рис. 4, см. маркеры) и определяем характер зависимости f_k от k. Так как линия тренда близка к прямой $f_k = ck$, принимаем гипотезу о пропорциональной зависимости f_k от k.

Шаг 6. Веса f_5, ..., f_{2000} (длина набора весов взята с небольшим по сравнению с длиной выборки $\hat{Q}_k$ запасом) задаем в виде $f_k = c_0 k$ и полагаем $f_{2001} = 0$. В первом приближении коэффициент c_0 принимаем равным единице.

Шаг 7. Полагая $a = m$, рассчитываем реализуемое по полученному набору весов распределение $\mathbf{Q}_1^{2001} = Q_1, \ldots, Q_{2001}$ (см. п. 1.2).

Шаг 8. Завершаем калибровку подбором c_0, при котором распределение $\mathbf{Q}_1^{2001}$ имеет среднее $\langle k \rangle = 2m$ (равенство $\langle f \rangle = a$ при таком c_0 выполняется автоматически). Получаем $c_0 = 0{,}8603$, и при $k \geqslant 5$ задаем веса общей формулой $f_k = 0{,}8603k$. На рис. 5 сравниваются заданное (маркеры) и реализуемое калиброванным генератором (сплошная линия) распределения.

Графики на рис. 5—7 построены в логарифмических шкалах.

Калиброванные генераторы выращивают графы с распределениями степени связности (РСС) (рис. 5—7, см. сплошные линии), хорошо согласующимися с эмпирическими вероятностями степеней (см. маркеры). Эти генераторы можно рассматривать и как результаты идентификации сетей

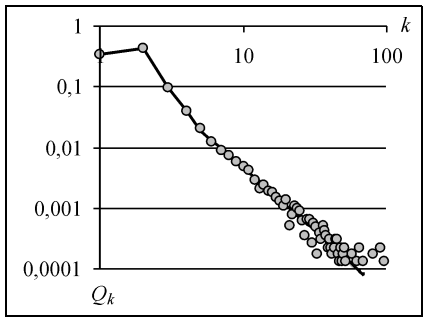

Рис. 5. Распределение степени связности калиброванного графа и сети автономных систем Интернет

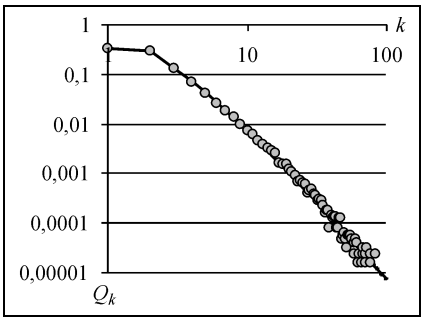

Рис. 6. Распределение степени связности калиброванного графа и сети маршрутизаторов Интернет

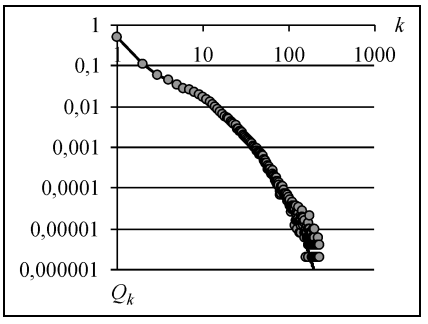

Рис. 7. Распределение степени связности калиброванного графа и сети участия актеров в общих фильмах

предпочтительного связывания. Рис. 6 характеризует качество идентификации сети маршрутизаторов Интернет по данным [9] (N = 124651, R = 207214). Параметры генератора: r_1 = 0,435, r_2 = 0,565, f_1 = = 0,272, f_2 = 1,224, f_k = 0,512ln(k) + 0,372k, $k \geqslant 3$. Рис. 7 характеризует идентификацию сети участия актеров в общих фильмах [10] (N = 511 416 — без изолированных узлов, R = 1 463 331). Два соответствующих актерам узла связаны ребром, если эти актеры снимались в одном фильме. Здесь r_1, ..., $r_8 > 0$, численно фиксированы f_1, ..., f_{10}, и при $k > 10$ общая формула весов имеет вид f_k = 4,429ln(k).

Калиброванные графовые модели сетей можно использовать для анализа и оптимизации различных сетевых процессов. Интернет — «сеть сетей», представляет собой иерархическую структуру, модулями которой являются автономные системы (АС). Каждая АС — это система IP-сетей (IP — Internet Protocol), имеющих свои уникальные IP-адреса, и маршрутизаторов (шлюзов), управляемую одним или несколькими операторами, которые осуществляют единую политику маршрутизации сообщений в пределах АС, независимую от остальной части Интернета. В состав АС может входить от единиц до нескольких тысяч компьютеров. Связь между АС осуществляют «граничные» шлюзы, реализующие основной протокол динамической маршрутизации в Интернете — «протокол граничного шлюза» (англ. Border Gateway Protocol, или BGP). Используя BGP, граничные шлюзы прокладывают маршруты не между отдельными маршрутизаторами, а между АС, рассматриваемыми как неделимые элементы сети, имеющие минимальный набор параметров, необходимых для маршрутизации. Этот набор включает в себя уникальный номер АС, число IP-сетей в АС, их адреса и внутренние расстояния до этих сетей от данного граничного шлюза. Граф сети автономных систем, выращиваемый калиброванным генератором, можно использовать для анализа ее связности при случайных отказах элементов (т. е. при потере связи с отдельными АС) и для расчета критической вероятности отказа, при достижении которой сеть распадается на несвязные компоненты. Администраторы АС, учитывая результаты такого анализа, могут уточнять свой выбор числа провайдеров и смежных АС для соединения с Интернетом. Уточнение выбора может привести к системному эффекту, несколько изменяющему ход развития Интернета. Этот эффект можно оценивать заранее, генерируя графы по первоначальным и по скорректированным правилам предпочтительного связывания, и используя в качестве затравки известное мгновенное состояние сети АС (например, опубликованное в работе [8]), близкое к текущему состоянию.

Калиброванный случайный граф, многократно реализуемый для определения средних значений искомых показателей, обладает рядом преимуществ по сравнению с графом известного мгновенного состояния сети. Поскольку большие сети (например, отдельные АС и Интернет в целом) быстро изменяются, сохраняя лишь свои существенные структурные свойства, то оптимизация решений под фиксированное мгновенное состояние сети может приводить при их реализации к результатам, далеким от ожидаемых. Отладку решений, тиражируемых для применения в разных вариантах системы (например, алгоритмов внутренней в пределах АС маршрутизации), также лучше осуществлять на нескольких реализациях калиброванного случайного графа, моделирующего систему. Кроме того, калиброванные графы позволяют моделировать развитие сетей в будущем.

Для эффективной отладки алгоритмов маршрутизации нужны гетерогенные калиброванные графы, содержащие вершины двух и более типов (для внутренней маршрутизации в пределах АС — вершины маршрутизаторов и IP-сетей). Это требует дальнейшего обобщения правила предпочтительного связывания в направлении использования приращений двух и более типов. Для выбора пути передачи сообщений алгоритмы маршрутизации используют метрики, учитывающие число промежуточных маршрутизаторов, задержки передачи, пропускную способность каналов и IP-сетей, денежную стоимость связи и т. д.; соответствующие параметры также следует правильно распределять по вершинам графа. Заметим, что подобные параметры каналов учитываются и владельцами новых АС при выборе узлов для подключения АС к Интернету. В этом состоит одна из возможных причин нелинейного характера зависимости веса f_k узла от его степени k в сети АС (для k = 1, ..., 4).

Графы социальных сетей, подобных сети участия актеров в общих фильмах, могут применяться в социальных исследованиях и проектах. Логарифмическая весовая функция в генераторе графа сети участия актеров, найденная при калибровке, может быть объяснена типичной для человека логарифмической шкалой восприятия уровня сигналов (в данном случае — актерской славы). Важное приложение найденных методов в социальной сфере заключается в исследовании сетей распространения инфекций и разработке на калиброванных моделях этих сетей надежных методов борьбы с эпидемиями. В целом область приложения разработанных методов не менее широка, чем у БА-графов, и наряду с Интернетом и социумом включает в себя сеть биохимических реакций в организме и разработку новых лекарств, пищевые цепи и экосистемы, энергетические и транспортные сети и др. [3].

4.5. Задача структурной идентификации сетей

Моделирование на калиброванных графах сетевых процессов (таких, как распространение инфекции при вакцинации узлов или формирование контактных кластеров при случайных отказах элементов сети) приводит к результатам, близким к результатам моделирования этих процессов непосредственно на исходных графах сетей. Однако при этом наблюдается небольшое, но устойчивое различие сравниваемых результатов, объясняемое тем, что выращиваемый граф, согласованный с исходной сетью по числу вершин и по распределению их степени связности, отличается от нее более тонкими структурными характеристиками. Например, коэффициент кластеризации $C = 3n_\Delta/n_V$ [11], определяемый отношением среднего числа n_Δ «треугольников» в графе к среднему числу n_V «вилок» (т. е. путей длины 2), у калиброванных по распределению степеней графов, представленных на рис. 5—7, получается в несколько раз меньшим, чем у исходных сетей. Обнаружение подобных различий приводит к задаче структурной идентификации сетей и структурной калибровки генераторов. А для этого необходимо глубже исследовать структуру графов предпочтительного связывания.

Одно из направлений такого исследования, развиваемое в работе [7], состоит в изучении вероятностных характеристик случайно выбранного ребра графа. Знание этих характеристик позволяет учитывать неравноценность вершин в слое, выбираемом для связывания, и регулировать структурные характеристики графа путем дополнительного сравнения и выбора вершин внутри слоя, как это делается, например, в предложенном в работе [7] алгоритме сепарабельной калибровки графа по коэффициенту кластеризации C. Применение данного алгоритма — например, в процессе выращивания графа, моделирующего сеть маршрутизаторов Интернет (см. рис. 6) — позволяет реализовать требуемое значение C и существенно сблизить результаты моделирования сетевых процессов на исходном графе сети и на его модели. В общем же случае, очевидно, наряду с коэффициентом кластеризации необходимо исследовать другие, более сложные индикаторы структурных особенностей графа.

ЗАКЛЮЧЕНИЕ

Нелинейное правило предпочтительного связывания расширяет возможности моделирования больших сетей, в которых вероятность выбора узла для связывания с ним не обязательно пропорциональна степени связности этого узла. Точные формулы распределения степени связности, найденные в статье для случайных графов с нелинейным правилом предпочтительного связывания и стохастическим приращением, включают в себя ранее установленные результаты в качестве частных случаев. Разработанные на основе этих формул методы калибровки генераторов просты в практическом применении и позволяют быстро выращивать графы, содержащие миллионы вершин и точно реализующие заданные распределения степени связности. Это создает предпосылки для эффективного решения задач структурной идентификации больших сетей методами аналитико-имитационного моделирования. Калиброванные случайные графы могут использоваться как для моделирования и оптимизации процессов, происходящих в исследуемых сетях, так и для тестирования разнообразных алгоритмов и программ, предназначенных для работы с графами.

ЛИТЕРАТУРА

1. *Erdös P., Rényi A.* On random graphs I. // Publ. Math. Debrecen. — 1959. — Vol. 6. — P. 290—297.
2. *Barabási, Albert-László and Albert, Réka.* Emergence of scaling in random networks // Science, 286:509-512, October 15, 1999.
3. *Барабаши А., Бонабо Э.* Безмасштабные сети // В мире науки. — 2003. — № 8. — С. 55—63.
4. *Barabási, Albert-László.* Scale-Free Networks: A Decade and Beyond // SCIENCE. — 2009. — Vol. 325. — P. 412—413.
5. *Dorogovtsev, S.N. and Mendes, J.F.F. and Samukhin, A.N.* Structure of Growing Networks: Exact Solution of the Barabasi-Albert's Model // Phys. Rev. Lett. 85, 4633 (2000).
6. *Guclu, H. and Yuksel, M.* Scale-Free Overlay Topologies with Hard Cutoffs for Unstructured Peer-to-Peer Networks // In Proc. of IEEE International Conference on Distributed Computing Systems (*ICDCS*), 2007.
7. *Задорожный В.Н., Юдин Е.Б.* Структурные свойства scale-free графа Барабаши-Альберт / Препринт материалов, принятых для публикации в журнале «Автоматика и телемеханика». 2009. — URL: http://zadorozhnyi.acouy.ru/articles/ZadorozhnyiYudinLastVersion.pdf (дата обращения: 01.03.2010).
8. *Структура* автономных систем сети Интернет, воссозданная на основе BGP таблиц, 2006 г. — URL: http://www-personal.umich.edu/~mejn/netdata/as-22july06.zip (дата обращения: 01.09.2009).
9. *Структура* сети маршрутизаторов Интернет (2006 г.). — URL: http://www.cise.ufl.edu/research/sparse/mat/Pajek/internet.mat (дата обращения: 01.09.2009).
10. *Структура* сети участия актеров в общих фильмах. — URL: http://www.nd.edu/~networks/resources/actor/actor.dat.gz (дата обращения: 03.02.2010).
11. *Newman M. E. J.* The structure and function of complex networks // SIAM Review 45, 167-256 (2003).

Статья представлена к публикации членом редколлегии Ф.Т. Алескеровым.

Задорожный Владимир Николаевич — канд. техн. наук, доцент, Омский государственный технический университет, ☎(3812) 65-20-84, ✉ zwn@yandex.ru.