Серия «Математика»
2018. Т. 25. С. 63—78

Онлайн-доступ к журналу:
http://mathizv.isu.ru

ИЗВЕСТИЯ
Иркутского
государственного
университета

УДК 519.176:519.178
MSC 05C70
DOI https://doi.org/10.26516/1997-7670.2018.25.63

Кластеризация графов на основе оценок изменения модулярности *

Н. Н. Мартынов
Бурятский государственный университет, Улан-Удэ, Российская Федерация

О. В. Хандарова
Институт монголоведения, буддологии и тибетологии СО РАН, Улан-Удэ, Российская Федерация

Ф. В. Хандаров
Бурятский государственный университет, Улан-Удэ, Российская Федерация

Аннотация. Кластеризация графов является одной из постоянно актуальных задач анализа данных. Существуют различные постановки данной задачи. Рассматривается задача поиска разбиения множества вершин на непересекающиеся подмножества таким образом, чтобы плотность связей между вершинами одного подмножества была выше, чем между вершинами различных подмножеств. Для решения данной задачи применяются различные подходы, многие из которых используют такую апостериорную оценку качества кластеризации, как модулярность. Функционал модулярности, принимая значение из [–1, 1], позволяет в формальном виде оценить качество кластеризации (разбиения на подмножества).

Предложен подход, позволяющий вместо расчета модулярности пользоваться менее вычислительно сложной оценкой ее изменения при операции объединения кластеров. Для разных видов графов сформулирован ряд теорем, показывающих возможность применения предлагаемой оценки вместо прямого вычисления модулярности.

Описана жадная алгоритмическая схема кластеризации, а также AMVE (Algorithm based on Modularity Variation Estimation) — простейший жадный алгоритм на ее основе. На тестовом проблемсете произведен сравнительный анализ AMVE с эвристическими алгоритмами кластеризации, реализованными в современных пакетах анализа графов: демонстрируется сравнительное преимущество AMVE как по скорости, так и по качеству кластеризации.

Также приводятся сведения об использовании разработанного программного обеспечения для анализа данных в социологии и литературоведении. В этих исследованиях рассматривались графы, построенные на данных социальных сетей (в качестве ребер использовалось отношение «дружбы» в социальной сети между

пользователями). Продемонстрировано небольшое превосходство AMVE по качеству кластеризации по сравнению с известными алгоритмами Louvain и Walktrap.

Ключевые слова: кластеризация графов, модулярность, выделение сообществ, анализ социальных сетей.

1. Введение

Экстремальные задачи теории графов, допускающие численное решение, возникают повсеместно в социологии, маркетинге, информационном поиске, логистике и иных областях. Одной из наиболее востребованных задач анализа графов является кластеризация множества вершин — выделение таких непересекающихся подмножеств, в каждом из которых вершины связаны между собой более чем с вершинами вне данного подмножества.

Кластеризация графов (выделение сообществ в графах) зачастую основывается на вычислении тех или иных критериев оптимальности разбиения графа на подмножества. Одной из таких наиболее широко используемых оценок качества кластеризации является *модулярность* [5;11]. Нахождение оптимального значения функционала модулярности является NP-полной задачей и получение точного ее решения перебором на рабочей станции достаточно затратно. Так, например, в инструкции к процедуре cluster_optimal для поиска точного решения в пакете iGraph для среды R отмечается, что графы до 50 вершин будут обрабатываться уверенно, до пары сотен вершин — также могут быть посильны [3]. Для практики такие размерности явно недостаточны, что влечет создание большого числа эвристических алгоритмов, использующих модулярность лишь в качестве промежуточной или итоговой оценки качества полученного разбиения. Такие алгоритмы могут как реализовывать общие комбинаторные эвристики (например, эволюционные алгоритмы [10;12]), так и основываться на более специфических характеристиках и процедурах обработки графов (например, механизмы поиска потоков в сетях [4] или анализ дендрограмм [14] и др. [5;8;9]). Вместе с тем, как очевидно, повысить эффективность с сохранением точности возможно при организации расчета альтернативных функционалу модулярности оценок, например, совпадающих с ним по знаку, или рассчитывающих лишь изменение функционала при перемещении вершин между кластерами. В настоящей работе описывается общая жадная алгоритмическая схема, эксплуатирующая гипотетические оценки изменения некоторого функционала качества; приводятся теоретически обоснованные оценки для функционала модулярности, пригодные

* Работа выполнена при финансовой поддержке РФФИ, гранты 17-06-00340 А, 18-312-00186 мол_а

для использования в описанной схеме; проводится эмпирический анализ производительности данного алгоритма в сравнении с рядом иных известных средств.

2. Постановка задачи кластеризации

Рассмотрим граф $G = \langle V, E \rangle$, где V — множество вершин, E— множество ребер. Пусть C — множество всех непустых подмножеств множества V. *Разбиением на кластеры* множества V вершин графа G будем называть такое отображение $\varphi : V \to C$, для которого выполняется $\mathbb{E}(\varphi) = \{C_j\}_{j=\overline{1,k}} \subset C$, $\forall i, j\, (1 \leq i, j \leq k) : C_i \cap C_j = \varnothing$, $V = \bigcup\limits_{i=1}^{k} C_i$.

Элементы множества значений $\mathbb{E}(\varphi)$ отображения φ будем называть *кластерами*. Множество всех φ, возможных для графа G, будем обозначать, как $\Phi = \Phi_G$. Пусть задана некоторая оценка разбиения $Q(\varphi) \in \mathbb{R}$, тогда задача кластеризации множества вершин V графа G может быть записана, как экстремальная:

$$\varphi* = \arg\max_{\varphi \in \Phi} Q(\varphi) \tag{2.1}$$

Для построения жадной алгоритмической схемы решения задачи 2.1, определяющей правила получения из текущего разбиения φ_k нового разбиения φ_{k+1} такого, что $Q(\varphi_k) \leq Q(\varphi_{k+1})$, будем в некотором порядке $P = \left(u_0, u_1, ..., u_{|V|}\right)$ перебирать вершины множества V. Для очередной вершины $u_i \in P$. станем поочередно рассматривать все кластеры $C_j \in \mathbb{E}(\varphi_k)$, кроме того, которому принадлежит в настоящий момент. Выберем тот из них, чтобы при перемещении в него получить новое разбиение , при котором $\Delta Q(\varphi_k, \varphi_{k+1}) = Q(\varphi_{k+1}) - Q(\varphi_k)$ было бы максимальным. Поскольку φ_{k+1} получено перемещением вершины в кластер , будем обозначать $\Delta Q(\varphi_k, \varphi_{k+1})$ далее, как $\Delta Q(u, C_j)$, и можем записать локальную задачу поиска кластера для перемещения в виде:

$$C_{local} = \arg\max_{C_j \in \mathbb{E}(\varphi_k)} \Delta Q(u, C_j) \tag{2.2}$$

Алгоритмическая схема решения задачи 2.1, таким образом, принимает следующий вид:

1) Для V определяем начальное разбиение $\varphi_0 = START(V)$.
2) Определяем порядок $P = P(V)$, $i = 0$, $k = 0$.
3) Пытаемся переместить $u_i \in P$ в кластер C_k, определяемый решением задачи 2.2 относительно u_i; если попытка перемещения удалась, то получаем новое разбиение φ_{k+1} и увеличиваем $k = k + 1$.

4) $i = i + 1$, если $i \leq |V|$ — шаг 3, иначе — шаг 5.

5) Если $k > 0$, то — шаг 2, иначе — шаг 6.

6) Выход.

В качестве оценки Q будем рассматривать модулярность, которая в текущих обозначениях может быть записана в виде 2.3 для невзвешенного графа, и в виде 2.4 для взвешенного.

$$Q\left(G,\varphi,\alpha\right)=\frac{1}{2\left|E\right|}\sum_{ij}\left(A_{ij}-\alpha\frac{d_i d_j}{2\left|E\right|}\right)\sigma\left(v_i,v_j\right), \tag{2.3}$$

$$Q_{ow}\left(G,\varphi,\alpha\right)=\frac{1}{2W}\sum_{ij}\left(A^w_{ij}-\alpha\frac{s_i s_j}{2W}\right)\sigma(v_i,v_j). \tag{2.4}$$

В случае, если граф ориентированный, формулы 2.3, 2.4 видоизменяются, соответственно, в виде 2.5, 2.6

$$Q\left(G,\varphi,\alpha\right)=\frac{1}{\left|E\right|}\sum_{ij}\left(A_{ij}-\alpha\frac{d^{out}_i d^{in}_j}{\left|E\right|}\right)\sigma\left(v_i,v_j\right), \tag{2.5}$$

$$Q_{ow}\left(G,\varphi,\alpha\right)=\frac{1}{W}\sum_{ij}\left(A^w_{ij}-\alpha\frac{s^{out}_i s^{in}_j}{W}\right)\sigma(v_i,v_j), \tag{2.6}$$

В формулах 2.3, 2.4, 2.5, 2.6 используются следующие обозначения: A^w_{ij} — вес ребра от вершины i к вершине j (суммарный вес в случае кратных ребер), A_{ij} — количество ребер от вершины i к вершине j, $|E|$ — мощность множества всех ребер графа, W — суммарный вес всех ребер графа, d^{out}_i и d^{in}_i — полустепени исхода и захода вершины i соответственно, s^{out}_i и s^{in}_i — суммарный вес всех исходящих и входящих ребер вершины i соответственно, $\sigma\left(v_i,v_j\right)=\begin{cases}1,\varphi(v_i)=\varphi(v_j)\\0,\varphi(v_i)\neq\varphi(v_j)\end{cases}$. Параметр α в формуле вводится для регулирования количества кластеров в получаемом разбиении и обычно берется из интервала $[0,2]$ [5;7;11]. Далее в описании алгоритма AMVE описан один из возможных способов обращения с данным параметром.

При такой оценке Q очевидно, что целевой кластер C_k на шаге 3 алгоритмической схемы имеет смысл искать лишь среди тех $C_j \in \mathbb{E}\left(\varphi_k\right)$, в которых есть вершины, смежные u_i. Обозначим множество таких кластеров, как $N\left(u_i,\varphi_k\right)\subseteq\mathbb{E}\left(\varphi_k\right)$, и станем называть его множеством кластеров, соседних вершине u_i.

Способы определения начального разбиения $\varphi_0 = START\left(V\right)$ и порядка перебора перемещаемых вершин $P = P\left(V\right)$, как очевидно, могут быть определены различными способами. В настоящей работе мы брали

φ_0 таким, чтобы каждая вершина попадала в отдельное подмножество, а в качестве $P = P(V)$ брали случайную перестановку множества V.

Наиболее вычислительно сложным этапом описанной алгоритмической схемы является пересчет оценки Q, необходимый для вычисления ΔQ, на шаге 3 при переборе возможных кластеров. Далее в настоящей работе описан способ расчета альтернативных оценок $\Delta\overline{Q}$, позволяющий значительно сократить эти вычисления.

3. Описание оценки $\Delta\overline{Q}$

Опишем в данном разделе вводимую оценку $\Delta\overline{Q}$, сформулировав ряд теорем для ориентированного и неориентированного, взвешенного и невзвешенного графов. Для этого уточним некоторые обозначения:

$\mathbb{E}\left(\varphi^{(t)}\right) = \left\{C_1^{(t)}, C_2^{(t)}, \ldots, C_h^{(t)}, \ldots, C_g^{(t)}, \ldots, C_k^{(t)}\right\}$ — множество кластеров для текущего разбиения $\varphi^{(t)}$,

$\mathbb{E}\left(\varphi^{(t+1)}\right) = \left\{C_1^{(t+1)}, C_2^{(t+1)}, \ldots, C_{gh}^{(t+1)}, \ldots, C_{k-1}^{(t+1)}\right\}$ — множество, полученное из $\mathbb{E}\left(\varphi^{(t)}\right)$ объединением пары кластеров $C_g^{(t+1)} = C_g^{(t)} \bigcup C_h^{(t)}$,

$D_i = \sum\limits_{j \in C_i} d_j$ — количество всех ребер, инцедентных вершине i,

$D_i^{out} = \sum\limits_{j \in C_i} d_j^{out}$ — количество всех ребер, исходящих из вершины i,

$D_i^{in} = \sum\limits_{j \in C_i} d_j^{in}$ — количество всех ребер, входящих в вершину i,

$S_i = \sum\limits_{j \in C_i} s_j$ — сумма весов всех ребер, инцедентных вершине i,

$S_i^{out} = \sum\limits_{j \in C_i} s_j^{out}$ — сумма весов всех ребер, исходящих из вершины i,

$S_i^{in} = \sum\limits_{j \in C_i} s_j^{in}$ — сумма весов ребер, входящих в вершину i,

$e(C_g, C_h) = \sum\limits_{\substack{i \in C_g \\ j \in C_h}} A_{ij}$ — количество ребер между C_g и C_h,

$e_o(C_g, C_h) = \sum\limits_{\substack{i \in C_g \\ j \in C_h}} (A_{ij} + A_{ji})$ — количество ребер из C_g в C_h,

$w(C_g, C_h) = \sum\limits_{\substack{i \in C_g \\ j \in C_h}} A_{ij}^w$— суммарный вес ребер между C_g и C_h,

$w_o(C_g, C_h) = \sum\limits_{\substack{i \in C_g \\ j \in C_h}} A_{ij}^w + A_{ji}^w$ — суммарный вес всех ребер из C_g в C_h.

Далее приводится полное доказательство теоремы 1, остальные теоремы доказываются похожим образом, поэтому их доказательства изложены менее подробно.

Теорема 1. *Пусть оценка Q разбиения невзвешенного неориентированного графа задана в виде 2.3, тогда при*

$$\Delta\overline{Q}\left(C_g, C_h, \alpha\right) = 2e(C_g, C_h) - \frac{\alpha}{|E|} D_g D_h > 0$$

выполняется $Q\left(G, \varphi^{(t)}, \alpha\right) < Q\left(G, \varphi^{(t+1)}, \alpha\right)$.

Доказательство. Благодаря функции σ в 2.3 в суммировании участвуют только те пары вершин, которые лежат в одном кластере, потому можно просуммировать Q по каждому кластеру отдельно:

$$Q = \frac{1}{2\,|E|} \sum_{h=1}^{k} \sum_{i,j \in C_h} \left(A_{ij} - \alpha \frac{d_i d_j}{2\,|E|} \right)$$

$$= \frac{1}{2\,|E|} \sum_{h=1}^{k} \left(\sum_{i,j \in C_h} A_{ij} - \frac{\alpha}{2\,|E|} \sum_{i,j \in C_h} d_i d_j \right),$$

где k — количество кластеров. Обозначив $L_h = \sum\limits_{i \in C_h} A_{ii}$ — суммарное количество петель у вершин в кластере C_h, получим

$$Q = \frac{1}{2\,|E|} \left(2 \sum_{h=1}^{k} |E_h| - \sum_{h=1}^{k} L_h - \frac{\alpha}{2\,|E|} \sum_{h=1}^{k} \sum_{i,j \in C_h} d_i d_j \right).$$

Отметим, что учетом петель в графе можно пренебречь, поскольку наличие петель у некоторой вершины не влияет на ее принадлежность к тому или иному кластеру. Обозначая $\sum\limits_{i,j \in C_h} d_i d_j = \sum\limits_{i \in C_h} d_i \left(\sum\limits_{j \in C_h} d_j \right) = \left(\sum\limits_{i \in C_h} d_i \right)^2 = {D_h}^2$, получим $Q = \frac{1}{2|E|} \left(2 \sum\limits_{i=1}^{k} |E_i| - \sum\limits_{i=1}^{k} L_i - \frac{\alpha}{2|E|} \sum\limits_{i=1}^{k} {D_i}^2 \right)$.

Поскольку множитель $\frac{1}{2|E|}$ является константой для одного и того же графа, то $\overline{Q} = 2\,|E|\,Q$ и $\arg\max \overline{Q} = \arg\max Q$.

Пусть $C_g = C_g^{(t)}, C_h = C_h^{(t)}, C_{gh} = C_g^{(t+1)} = C_g \bigcup C_h$, тогда имея $|E_{gh}| = |E_g| + |E_h| + e(C_g, C_h)$, ${D_{gh}}^2 = {D_g}^2 + {D_h}^2 + 2 D_g D_h$ и $L_{gh} = L_g + L_h$, и выполняя $\Delta\overline{Q}\left(C_g, C_h, \alpha\right) = \overline{Q}\left(G, \hat{\varphi}, \alpha\right) - \overline{Q}\left(G, \varphi, \alpha\right)$, получим: $\Delta\overline{Q}\left(C_g, C_h, \alpha\right) = 2e(C_g, C_h) - \frac{\alpha}{|E|} D_g D_h$.

Таким образом, при $\Delta\overline{Q}\left(C_g, C_h, \alpha\right) > 0$ получаем $\overline{Q}\left(G, \varphi^{(t)}, \alpha\right) < \overline{Q}\left(G, \varphi^{(t+1)}, \alpha\right)$ и, соответственно, $Q\left(G, \varphi^{(t)}, \alpha\right) < Q\left(G, \varphi^{(t+1)}, \alpha\right)$. Теорема доказана.

□

Теорема 2. *Пусть оценка Q_w разбиения взвешенного неориентированного графа задается в виде 2.4, тогда при*

$$\Delta\overline{Q}_w\left(C_g, C_h, \alpha\right) = 2w\left(C_g, C_h\right) - \frac{\alpha}{W} S_g S_h > 0$$

выполняется $Q_w\left(G, \varphi^{(t)}, \alpha\right) < Q_w\left(G, \varphi^{(t+1)}, \alpha\right)$.

Доказательство. Просуммируем Q_w по каждому кластеру отдельно

$$\begin{aligned} Q_w &= \frac{1}{2W} \sum_{h=1}^{k} \sum_{i,j \in C_h} \left(A_{ij}^w - \alpha \frac{s_i s_j}{2W} \right) \\ &= \frac{1}{2W} \sum_{h=1}^{k} \left(\sum_{i,j \in C_h} A_{ij}^w - \frac{\alpha}{2W} \sum_{i,j \in C_h} s_i s_j \right) = \\ &= \frac{1}{2W} \left(2 \sum_{h=1}^{k} W_h - \sum_{h=1}^{k} L_h^w - \frac{\alpha}{2W} \sum_{h=1}^{k} \sum_{i,j \in C_h} s_i s_j \right), \end{aligned}$$

где $L_h^w = \sum\limits_{i \in C_h} A_{ii}^w$, $W_h = \sum\limits_{i,j \in V_i} A_{ij}^w$.

Выражая $\sum\limits_{i,j \in C_h} s_i s_j = \sum\limits_{i \in C_h} s_i \left(\sum\limits_{j \in C_h} s_j \right) = \left(\sum\limits_{i \in C_h} s_i \right)^2 = {S_h}^2$, получим $Q_w = \frac{1}{2W} \left(2 \sum\limits_{i=1}^{k} W_i - \sum\limits_{i=1}^{k} L_i^w - \frac{\alpha}{2W} \sum\limits_{i=1}^{k} {S_i}^2 \right)$ и $\overline{Q}_w = 2 \sum\limits_{i=1}^{k} W_i - \sum\limits_{i=1}^{k} L_i^w - \frac{\alpha}{2W} \sum\limits_{i=1}^{k} {S_i}^2$. С учетом $W_{gh} = W_g + W_h + w(C_g, C_h)$, ${S_{gh}}^2 = {S_g}^2 + {S_h}^2 + 2 S_g S_h$, $L_{gh}^w = L_g^w + L_h^w$, получим

$$\Delta\overline{Q}_w\left(C_g, C_h, \alpha\right) = 2w\left(C_g, C_h\right) - \frac{\alpha}{W} S_g S_h.$$

Таким образом, при $\Delta\overline{Q}_w\left(C_g, C_h, \alpha\right) > 0$ получаем $\overline{Q}_w\left(G, \varphi^{(t)}, \alpha\right) < \overline{Q}_w\left(G, \varphi^{(t+1)}, \alpha\right)$ и, соответственно, $Q_w\left(G, \varphi^{(t)}, \alpha\right) < Q_w\left(G, \varphi^{(t+1)}, \alpha\right)$. Теорема доказана. □

Теорема 3. *Пусть оценка Q_o разбиения невзвешенного ориентированного графа задана в виде 2.5, тогда при*

$$\Delta\overline{Q}_o\left(C_g, C_h, \alpha\right) = e_o\left(C_g, C_h\right) - \frac{\alpha}{|E|} \left(D_h^{out} D_g^{in} + D_g^{out} D_h^{in} \right) > 0$$

выполняется $Q_o\left(G, \varphi^{(t)}, \alpha\right) < Q_o\left(G, \varphi^{(t+1)}, \alpha\right)$.

Доказательство. Просуммируем Q_o по каждому кластеру отдельно

$$Q_o = \frac{1}{|E|}\sum_{h=1}^{k}\sum_{i,j\in C_h}\left(A_{ij} - \alpha\frac{d_i^{out}d_j^{in}}{|E|}\right) =$$

$$= \frac{1}{|E|}\sum_{h=1}^{k}\left(\sum_{i,j\in C_h} A_{ij} - \frac{\alpha}{|E|}\sum_{i,j\in C_h} d_i^{out}d_j^{in}\right) =$$

$$= \frac{1}{|E|}\left(\sum_{h=1}^{k}|E_h| - \frac{\alpha}{|E|}\sum_{h=1}^{k}\sum_{i,j\in C_h} d_i^{out}d_j^{in}\right).$$

Выражая $\sum\limits_{i,j\in C_h} d_i^{out}d_j^{in} = \left(\sum\limits_{i\in C_h} d_i^{out}\right)\left(\sum\limits_{i\in C_h} d_i^{in}\right) = D_h^{out}D_h^{in}$, получим $Q_o = \frac{1}{|E|}\left(\sum\limits_{h=1}^{k}|E_h| - \frac{\alpha}{|E|}\sum\limits_{h=1}^{k} D_h^{out}D_h^{in}\right)$ и $\overline{Q}_o = \sum\limits_{h=1}^{k}|E_h| - \frac{\alpha}{|E|}\sum\limits_{h=1}^{k} D_h^{out}D_h^{in}$. С учетом $|E_{gh}| = |E_g| + |E_h| + e_o(C_g, C_h)$, $D_{gh}^{out}D_{gh}^{in} = D_g^{out}D_g^{in} + D_h^{out}D_h^{in} + \left(D_h^{out}D_g^{in} + D_g^{out}D_h^{in}\right)$ получим

$$\Delta\overline{Q}_o\left(C_g, C_h, \alpha\right) = e_o\left(C_g, C_h\right) - \frac{\alpha}{|E|}\left(D_h^{out}D_g^{in} + D_g^{out}D_h^{in}\right).$$

Таким образом, при $\Delta\overline{Q}_o\left(C_g, C_h, \alpha\right) > 0$ получаем

$$\overline{Q}_o\left(G, \varphi^{(t)}, \alpha\right) < \overline{Q}_o\left(G, \varphi^{(t+1)}, \alpha\right)$$

и, соответственно, $Q_o\left(G, \varphi^{(t)}, \alpha\right) < Q_o\left(G, \varphi^{(t+1)}, \alpha\right)$. □

Теорема 4. *Пусть оценка Q_{ow} разбиения взвешенного ориентированного графа задана в виде 2.6, тогда при*

$$\Delta\overline{Q}_{ow}\left(C_g, C_h, \alpha\right) = w_o\left(C_g, C_h\right) - \frac{\alpha}{W}\left(S_h^{out}S_g^{in} + S_g^{out}S_h^{in}\right) > 0$$

выполняется $Q_{ow}\left(G, \varphi^{(t)}, \alpha\right) < Q_{ow}\left(G, \varphi^{(t+1)}, \alpha\right)$.

Доказательство. Просуммируем Q_{ow} по каждому кластеру отдельно

$$Q_{ow} = \frac{1}{W}\sum_{h=1}^{k}\sum_{i,j\in C_h}\left(A_{ij}^{w} - \alpha\frac{s_i^{out}s_j^{in}}{W}\right) =$$

$$= \frac{1}{W}\sum_{h=1}^{k}\left(\sum_{i,j\in C_h} A_{ij}^{w} - \frac{\alpha}{W}\sum_{i,j\in C_h} s_i^{out}s_j^{in}\right) =$$

$$= Q_{ow} = \frac{1}{W}\left(\sum_{h=1}^{k} W_h - \frac{\alpha}{W}\sum_{h=1}^{k}\sum_{i,j\in C_h} s_i^{out}s_j^{in}\right),$$

Выражая $\sum\limits_{i,j\in C_h} s_i^{out} s_j^{in} = \left(\sum\limits_{i\in C_h} s_i^{out}\right)\left(\sum\limits_{i\in C_h} s_i^{in}\right) = S_h^{out} S_h^{in}$, получим $Q_{ow} = \frac{1}{W}\left(\sum\limits_{h=1}^{k} W_h - \frac{\alpha}{W}\sum\limits_{h=1}^{k} S_h^{out} S_h^{in}\right)$ и $\overline{Q}_{ow} = \sum\limits_{h=0}^{k} W_h - \frac{\alpha}{W}\sum\limits_{h=0}^{k} S_h^{out} S_h^{in}$. С учетом $W_{gh} = W_g + W_h + w_o(C_g, C_h)$ и $S_{gh}^{out} S_{gh}^{in} = S_g^{out} S_g^{in} + S_h^{out} S_h^{in} + \left(S_h^{out} S_g^{in} + S_g^{out} S_h^{in}\right)$ получим

$$\Delta\overline{Q}_{ow}\left(C_g, C_h, \alpha\right) = w_o\left(C_g, C_h\right) - \frac{\alpha}{W}\left(S_h^{out} S_g^{in} + S_g^{out} S_h^{in}\right).$$

Таким образом, при $\Delta\overline{Q}_{ow}\left(C_g, C_h, \alpha\right) > 0$ получаем

$$\overline{Q}_{ow}\left(G, \varphi^{(t)}, \alpha\right) < \overline{Q}_{ow}\left(G, \varphi^{(t+1)}, \alpha\right)$$

и, соответственно, $Q_o w\left(G, \varphi^{(t)}, \alpha\right) < Q_o w\left(G, \varphi^{(t+1)}, \alpha\right)$. □

4. Конкретизация алгоритма

Зададим некоторую процедуру получения нового графа G' как «сжатие» графа G согласно некоторому разбиению φ:

$$G' = COLLAPSE\left(G, \varphi\right).$$

Будем считать, что в данной процедуре каждому кластеру разметки φ графа G соответствует единственная вершина в графе G'. Ребрам внутри кластеров графа G, соответствуют кратные петли в графе G', а ребрам между вершинами из разных кластеров графа G соответствуют кратные ребра между вершинами графа G'. замечание.

Замечание 1. Кратные ребра между каждой парой вершин G' (с учетом ориентации ребер в ориентированных графах) можно заменить на единственное ребро, вес которого равен сумме весов этих кратных ребер.

Константу α из формулы вычисления модулярности, можно оставить фиксированной или предложить какой-либо способ её адаптации. В настоящей работе для α запущен простой перебор по сетке значений $[0.8, 1.2]$ с шагом 0.2.

Зададим некоторую метрику q, упрощающую работу с $\Delta\overline{Q}$, как:

$$q\left(u_i, C_j, \alpha\right) = \begin{cases} \Delta\overline{Q}\left(u_i, C_j, \alpha\right), & \text{если } u_i \notin C_j \\ \Delta\overline{Q}\left(u_i, C_j \backslash u_i, \alpha\right), & \text{если } u_i \in C_j \end{cases}.$$

Конкретизируем алгоритмическую схему с учетом вышесказанного.

1) Задаем граф $G^{(0)} = \left\langle V^{(0)}, E^{(0)} \right\rangle$, $t = 0$.

2) Строим граф $G^{(t+1)}$ на основе графа $G^{(t)}$

   а) Задаем начальное разбиение $\varphi_0^{(t)} = START\left(V^{(t)}\right)$ и $k = 0$

   б) Задаем $\alpha = \alpha_0$.

   в) Определяем порядок обхода вершин $P^{(t)} = P\left(V^{(t)}\right)$, $i = 0$.

   г) Пытаемся переместить $u_i^{(t)} \in P^{(t)}$:

   – Перебором находим $C_k^{(t)} = \arg \max\limits_{C_j \in N\left(u_i^{(t)}, \varphi_k^{(t)}\right)} q\left(u_i^{(t)}, C_j, \alpha\right)$.

   – Если $q\left(u_i^{(t)}, C_k^{(t)}, \alpha\right) \neq 0$, строим новое разбиение $\varphi_{k+1}^{(t)}$:

   – Если $q\left(u_i^{(t)}, C_k^{(t)}, \alpha\right) > 0$, перемещаем $u_i^{(t)}$ в существующий кластер $C_k^{(t)}$, и $k = k + 1$;

   – Если $q\left(u_i^{(t)}, C_k^{(t)}, \alpha\right) < 0$, создаём новый кластер $\widehat{C}_k^{(t)}$ и перемещаем в него $u_i^{(t)}$, и $k = k + 1$.

   – $i = i + 1$, если $i \leq \left|V^{(t)}\right|$ — шаг 2.4.1, иначе — шаг 2.5.

   д) Если $k > 0$ (перемещения на шаге 2.4 были), то — шаг 2.3, иначе — шаг 2.6.

   е) $\alpha = \alpha + step_\alpha$, если $\alpha > \alpha_{\max}$, то шаг 2.7, иначе - шаг 2.2.

   ж) «Сжимаем» граф $G^{(t)}$: $G^{(t+1)} = COLLAPSE\left(G^{(t)}, \varphi_k^{(t)}\right)$.

3) Если $\left|V^{(t+1)}\right| = \left|V^{(t)}\right|$ — шаг 4, иначе — $t = t + 1$ и — шаг 2.

4) Выход.

Описанный алгоритм, использующий $\Delta\overline{Q}$, позволяет повысить эффективность определения целевого кластера для перемещения в него вершины по сравнению с полным расчетом модулярности Q с $O\left(dn^2\right)$ до $O\left(d \log d\right)$, где d — степень перемещаемой вершины, n — число вершин в графе. Далее в статье будем обозначать разработанный алгоритм, как AMVE (Algorithm based on Modularity Variation Estimation).

5. Сравнительный анализ

AMVE сравнивался с широко употребляемыми алгоритмами Louvain [2], Walktrap [13] и Label Propagation [15], для которых использовалась реализация из библиотеки iGraph пакета R [3]. AMVE реализован в виде программного комплекса на C++.

Тестирование производилось на наборе из 100 неориентированных невзвешенных графов, сгенерированных так, как описано далее. После задания случайного числа вершин $200 \leq n \leq 1000$ графа и случайного разбиения множества вершин на подмножества размерности $3 \leq c \leq \lfloor n/4 \rfloor$, генерировались ребра с вероятностью $p \in [0.7, 0.9]$ — между каждой парой вершин внутри одного кластера и $p \in [0.1, 0.3]$ — между каждой парой вершин из разных кластеров. Таким образом, для каждого графа была известна «базовая» структура кластеров (в общем случае — неоптимальная, но, в силу способа генерации, уже достаточно приемлемая) и соответствующее «базовое» значение модулярности.

В таблице 1 показаны результаты кластеризации тестовых примеров. Видно, что AMVE по достигаемому значению Q (качеству кластеризации) немного превосходит Walktrap и Louvain, тогда как Label Propagation показывает результаты значительно хуже. Измерений точного времени работы не проводилось, т.к. алгоритмы реализованы в составе пакетных библиотек, но, можно утверждать, что AMVE и Louvain требуется примерно одинаковое время на кластеризацию, немного медленнее работает Walktrap, и с еще большим отставанием Label Propagation.

Таблица 1

Результаты сравнительного тестирования алгоритмов кластеризации

Алгоритмы кластеризации	AMVE	LP	Louvain	Walktrap
Количество примеров, на которых показан результат лучше конкурентов	41	0	38	20
Количество примеров, на которых достигнутое значение Q больше или равно «базового»	70	1	54	42

Дальнейшие улучшения алгоритма могут касаться способов генерации начального разбиения, порядка обхода вершин, способа адаптации параметра α. Как очевидно, алгоритм может быть достаточно эффективно распараллелен (в частности параллельная реализация напрашивается на шаге 2.4.1 — при независимом переборе целевых кластеров для перемещения вершины). Определенный интерес представляет также использование предложенных оценок в эволюционных схемах решения экстремальных задач (например, в операторах скрещивания и/или мутации в генетическом алгоритме). В любом случае перспективность подхода кажется достаточно явной в свете результатов тестирования описанной простейшей реализации AMVE.

6. Решение прикладных задач

Разработанное программное обеспечение использовалось для анализа данных в социологии и литературоведении. Исследовались графы, построенные на данных социальных сетей (в качестве ребер использовалось отношение «дружбы» в социальной сети между пользователями).

В таблице 2 приведены значение модулярности / количество кластеров при кластеризации с помощью алгоритмов AMVE, Walktrap и Louvain.

Таблица 2

Результаты работы алгоритмов кластеризации

Граф		AMVE	Walktrap	Louvain
Писатели Бурятии в сети Facebook	$\|V\| = 49$, $\|E\| = 196$	0.2/15	0.119/22	0.189/16
Редакторы российских литературных журналов	$\|V\| = 71$, $\|E\| = 604$	0.223/5	0.146/8	0.218/4
Буддисты соцсети VK	$\|V\| = 84927$, $\|E\| = 370875$	0.746/ 4570	0.664/ 12190	0.738/ 4572
Подписчики паблика соцсети VK «Хамбо-лама Д. Д. Итигэлов» [1]	$\|V\| = 24190$ $\|E\| = 85539$	0.721/ 296	0.632/ 1492	0.709 293

На графе писателей Бурятии в Facebook AMVE достигает глобального максимума модулярности, что улучшает результаты работы [6], в которой используется версия алгоритма с фиксированным $\alpha = 1$.

AMVE и Louvain демонстрируют похожую производительность с небольшим преимуществом AMVE: значения модулярности различаются на сотые или тысячные доли, количества выявленных кластеров отличается на несколько единиц даже на графе порядка $8 * 10^5$ вершин.

7. Заключение

В работе описана жадная алгоритмическая схема кластеризации графов, основанная на использовании модулярности — известного апостериорного критерия качества кластеризации. Предложен и теоретически обоснован эффективный способ расчета оценок изменения модулярности при объединении кластеров. Описан AMVE (Algorithm based on Modularity Variation Estimation) — простейший алгоритм на основе предложенных алгоритмической схемы и способа расчета оценок. Проведен сравнительный анализ AMVE с рядом используемых на практике алгоритмов кластеризации, показано преимущество AMVE, как при

тестировании, так и при решении прикладных задач. Намечены также пути дальнейшего развития подхода.

Список литературы

1. Бадмацыренов Т. Б., Скворцов М. В., Хандаров Ф. В. Буддийские цифровые практики трансцендентности: VK-сообщество "Хамбо Лама Даши-Доржо Итигэлов"// Мониторинг общественного мнения: экономические и социальные перемены. 2018. № 2. С. 309–332. https://doi.org/10.14515/monitoring.2018.2.18
2. Писатели Бурятии в сети Facebook: литературные репутации в виртуальном пространстве / О. В. Хандарова, Ф. В. Хандаров, Н. Н. Мартынов, М. В. Скворцов // Медиафилософия. 2017. Т. 13. С. 212-226.
3. Fast unfolding of communities in large networks / V. D. Blondel [et al.] //Journal of statistical mechanics: theory and experiment. 2008. Vol. 2008, N 10. P. 10008. https://doi.org/10.1088/1742-5468/2008/10/P10008
4. Csardi G., Nepusz T. The igraph software package for complex network research //InterJournal, Complex Systems. 2006. Vol. 1695, N 5. P. 1-9.
5. Flake G. W., Lawrence S., Giles C. L. Efficient identification of web communities //Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2000. P. 150-160. https://doi.org/10.1145/347090.347121
6. Fortunato S. Community detection in graphs //Physics reports. 2010. Vol. 486, N. 3-5. P. 75-174. https://doi.org/10.1016/j.physrep.2009.11.002
7. Le Martelot E., Hankin C. Fast multi-scale detection of relevant communities in large-scale networks //The Computer Journal. 2013. Vol. 56, N 9. P. 1136-1150. https://doi.org/10.1093/comjnl/bxt002
8. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters / J. Leskovec, K. J. Lang, A. Dasgupta, M. W. Mahoney //Internet Mathematics. 2009. Vol. 6. N 1. P. 29-123. https://doi.org/10.1080/15427951.2009.10129177
9. Leskovec J., Lang K. J., Mahoney M. Empirical comparison of algorithms for network community detection //Proceedings of the 19th international conference on World wide web. ACM, 2010. P. 631-640. https://doi.org/10.1145/1772690.1772755
10. Maulik U., Bandyopadhyay S. Genetic algorithm-based clustering technique //Pattern recognition. 2000. Vol. 33, N 9. P. 1455-1465. https://doi.org/10.1016/S0031-3203(99)00137-5
11. Newman M. E. J. Detecting community structure in networks //The European Physical Journal B. 2004. Vol. 38, N 2. P. 321-330. https://doi.org/10.1140/epjb/e2004-00124-y
12. Pizzuti C. A multiobjective genetic algorithm to find communities in complex networks // IEEE Transactions on Evolutionary Computation, 2012. Vol. 16, N 3. P. 418-430. https://doi.org/10.1109/TEVC.2011.2161090
13. Pons P., Latapy M. Computing communities in large networks using random walks // Journal of Graph Algorithms and Applications, 2006. Vol. 10, N 2. P. 191–218. https://doi.org/10.1007/11569596_31
14. Defining and identifying communities in networks / F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi //Proceedings of the National Academy of Sciences. 2004. Vol. 101, N 9. P. 2658-2663. https://doi.org/10.1073/pnas.0400054101

15. Raghavan U. N., Albert R., Kumara S. Near linear time algorithm to detect community structures in large-scale networks //Physical review E. 2007. Vol. 76. N 3. P. 036106. https://doi.org/10.1103/PhysRevE.76.036106
16. Reichardt J., Bornholdt S. Statistical mechanics of community detection //Physical Review E. 2006. Vol. 74, N 1. P. 016110. https://doi.org/10.1103/PhysRevE.74.016110

Никита Николаевич Мартынов, магистрант, Институт математики и информатики, Бурятский государственный университет, Российская Федерация, 670000, Республика Бурятия, г. Улан-Удэ, ул. Смолина, 24а, тел.: (+7914)6368355 (e-mail: supercell666@mail.ru)

Ольга Владимировна Хандарова, кандидат филологических наук, младший научный сотрудник, Институт монголоведения, буддологии и тибетологии СО РАН, Российская Федерация, 670047, Республика Бурятия, Улан-Удэ, ул. Сахьяновой, 6, тел.: (+7924)7731409 (e-mail: olga.khandarova@gmail.com)

Фёдор Владимирович Хандаров, кандидат технических наук, старший преподаватель,Институт математики и информатики, Бурятский государственный университет, Российская Федерация, 670000, Республика Бурятия, г. Улан-Удэ, ул. Смолина, 24а, тел.: (+7924)4563112 (e-mail: fedor.khandarov@gmail.com)

Поступила в редакцию 08.08.18

Graph Clustering Based on Modularity Variation Estimations

N. N. Martynov

Buryat State University, Ulan-Ude, Russian Federation

O. V. Khandarova

Institute for Mongolian, Buddhist and Tibetan Studies SB RAS, Ulan-Ude, Russian Federation

F. V. Khandarov

Buryat State University, Ulan-Ude, Russian Federation

Abstract. Graph clustering is one of the constantly actual data analysis problems. There are various statements of this problem. In this paper we consider the statement of search for a partition of a graph vertices set into disjoint subsets. In such a way, that the density of connections between the vertices of one subset was higher than that between the vertices of different subsets.

There is a lot of various approaches, and many of them use such as an a posteriori estimate of clustering quality, as modularity. The modularity functional, taking the value from [-1, 1], allows us to formally evaluate the quality of partitioning into subsets. This paper deals with an approach, instead of calculating the modularity, using a less

computationally complex estimate of modularity change in the operation of clusters union.

Four theorems for different graph types are formulated, presenting the possibility of application of suggested estimate, instead of direct modularity calculations. A greedy algorithmic scheme and also AMVE (Algorithm based on Modularity Variation Estimation) — simple greedy algorithm based on the scheme are proposed.

An attempt of comparative analysis on the test problemet of AMVE with heuristic clustering algorithms implemented in actual data analysis software is described. It is shown the comparative advantage of AMVE, both in terms of speed and quality of clustering.

Also, the cases on the use of developed algorithm and its implementation for data analysis in sociology and history and criticism of literature are mentioned. In these cases, investigated graphs based on social networks data (the ratio of "friendship" in the social network between users used as the graph edges). It is demonstrated a slight superiority of AMVE in clustering quality compared to the known algorithms Louvain and Walktrap.

Keywords: graph clustering, modularity, community detection, social network analysis.

References

1. Badmatsyrenov T.B., Skvortsov M.V., Khandarov F.V. Buddhist digital practices of transcendence: VK-community "Hambo Lama Dashi-Dorzho Itigilov". *Monitoring of Public Opinion: Economic and Social Changes*, 2018, no. 2, pp. 309-332. (in Russian) https://doi.org/10.14515/monitoring.2018.2.18.
2. Blondel V. D. et al. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008, vol. 2008, no. 10, pp. 10008. https://doi.org/10.1088/1742-5468/2008/10/P10008
3. Csardi G., Nepusz T. The igraph software package for complex network research. *InterJournal, Complex Systems*, 2006, vol. 1695, no. 5, pp. 1-9.
4. Flake G. W., Lawrence S., Giles C. L. Efficient identification of web communities. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000, pp. 150-160. https://doi.org/10.1145/347090.347121
5. Fortunato S. Community detection in graphs. *Physics reports*, 2010, vol. 486. N. 3-5, pp. 75-174. https://doi.org/10.1016/j.physrep.2009.11.002
6. Khandarova O.V., Khandarov F.V., Martynov N.N., Skvortsov M.V. Writers of Buryatia on Facebook: literary reputations in the virtual space *Mediafilosofiya*, 2017, vol. 13, pp. 212-226. (in Russian)
7. Le Martelot E., Hankin C. Fast multi-scale detection of relevant communities in large-scale networks. *The Computer Journal*, 2013, vol. 56, no. 9, pp. 1136-1150. https://doi.org/10.1093/comjnl/bxt002
8. Leskovec J., Lang K. J., Dasgupta A., Mahoney M. W. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 2009, vol. 6, no. 1, pp. 29-123. https://doi.org/10.1080/15427951.2009.10129177
9. Leskovec J., Lang K. J., Mahoney M. Empirical comparison of algorithms for network community detection. *Proceedings of the 19th international conference on World wide web. ACM*, 2010, pp. 631-640. https://doi.org/10.1145/1772690.1772755

10. Maulik U., Bandyopadhyay S. Genetic algorithm-based clustering technique. *Pattern recognition*, 2000, vol. 33, no. 9, pp. 1455-1465. https://doi.org/10.1016/S0031-3203(99)00137-5
11. Newman M. E. J. Detecting community structure in networks. *The European Physical Journal B*, 2004, vol. 38, no. 2, pp. 321-330. https://doi.org/10.1140/epjb/e2004-00124-y
12. Pizzuti C. A multiobjective genetic algorithm to find communities in complex networks. *IEEE Transactions on Evolutionary Computation*, 2012, vol. 16, no. 3, pp. 418-430. https://doi.org/10.1109/TEVC.2011.2161090
13. Pons P., Latapy M. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 2006, vol. 10, no. 2, pp. 191–218. https://doi.org/10.1007/11569596_31
14. Radicchi F., Castellano C., Cecconi F., Loreto V., Parisi D. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, 2004, vol. 101, no. 9, pp. 2658-2663. https://doi.org/10.1073/pnas.0400054101
15. Raghavan U. N., Albert R., Kumara S. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 2007, vol. 76, no. 3, pp. 036106. https://doi.org/10.1103/PhysRevE.76.036106
16. Reichardt J., Bornholdt S. Statistical mechanics of community detection. *Physical Review E*, 2006, vol. 74, no. 1, pp. 016110. https://doi.org/10.1103/PhysRevE.74.016110

Nikita Martynov, Undergraduate, Buryat State University, 24a, Smolin st., Ulan-Ude, Republic of Buryati, 670000, Russian Federation, tel.: (+7914)6368355 (e-mail: supercell666@mail.ru)

Olga Khandarova, Candidate of Sciences (Philology), Junior Research Scientist, Institute for Mongolian, Buddhist and Tibetan Studies SB RAS, 6, Sakhyanova st., Ulan-Ude, Republic of Buryatia, 670000, Russian Federation, tel.: (+7924)7731409 (e-mail: olga.khandarova@gmail.com)

Fedor Khandarov, Candidate of Sciences (Technical), Senior Lecturer, Buryat State University, 24a, Smolin st., Ulan-Ude, Republic of Buryatia, 670000, Russian Federation, tel.: (+7924)4563112 (e-mail: fedor.khandarov@gmail.com)

Received 08.08.18