УДК 519.178
MSC 05C70

Вестник СПбГУ. Прикладная математика. Информатика... 2018. Т. 14. Вып. 3

Метод максимального правдоподобия для выделения сообществ в коммуникационных сетях*

В. В. Мазалов[1,2], *Н. Н. Никитина*[2]

[1] Санкт-Петербургский государственный университет, Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

[2] Федеральный исследовательский центр «Карельский научный центр Российской академии наук», Российская Федерация, 185910, Петрозаводск, ул. Пушкинская, 11

Для цитирования: *Мазалов В. В., Никитина Н. Н.* Метод максимального правдоподобия для выделения сообществ в коммуникационных сетях // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2018. Т. 14. Вып. 3. С. 200–214. https://doi.org/10.21638/11702/spbu10.2018.302

Выделение сообществ в социальных и коммуникационных сетях является важной задачей во многих прикладных областях: биологии, социологии, социальных сетях, особенно актуально для тех сетей, которые представлены графами большой размерности. При этом важно использовать приближенные методы, которые позволяют за ограниченное время приводить, возможно, не к оптимальному результату, а к близкому к оптимальному. Предлагается метод выделения структуры сообществ на основе метода максимального правдоподобия. Описан алгоритм поиска структуры сообществ и проиллюстрирована работа алгоритма на численных примерах.

Ключевые слова: сетевые сообщества, выделение сообществ в сети, метод максимального правдоподобия, сэмплирование по Гиббсу.

1. Введение. Социальные сети играют важную роль в современном обществе. Формально их можно представить в виде коммуникационного графа, где вершинами являются члены социальной сети, а связи между ними — их переписка, общие интересы и общие друзья. В этом случае их можно анализировать с помощью математических методов. Социальные сети быстро изменяются, поэтому одним из инструментов их изучения служат случайные графы. Математический анализ графов социальных сетей производится с помощью различных мер центральности вершин и ребер графа. Одним из первых критериев центральности было понятие центральности по посредничеству (*betweenness centrality*) [1]. Популярен также метод PageRank, основанный на аналогии со случайными блужданиями [2–4]. Эффективным оказался метод, построенный на законах Кирхгоффа для электрических цепей [5–7]. В последнее время появились теоретико-игровые меры центральности [8, 9].

Реальные социальные сети характеризуются значительными нерегулярностями, зависящими от организации сетей [10, 11]. Это связано с тем, что в структуре коммуникационного графа есть группы вершин, для которых свойственно большее распределение связей внутри группы, чем с вершинами других групп. Учитывая такие особенности, можно выделить сообщества в сети. Это важно во многих исследованиях в биологии [11, 12], социологии [13], социальных сетях [14–16] и др. Выделение сообществ в социальной сети может помочь выявить аномальное поведение его

* Работа выполнена при финансовой поддержке Российского научного фонда (проект №17-11-01079).

участников. Вообще говоря, в выделении сообществ в сети есть два подхода, когда сообщества могут быть непересекающимися и когда они могут пересекаться. Последнее может произойти, если нас интересуют профессиональные, семейные, дружеские связи участников.

Один из первых подходов к кластеризации графов был предложен в работе [11]. Идея этого метода заключалась в последовательном отсеивании вершин графа с наибольшей мерой центральности и последующем отсечении иерархического дерева (дендрограммы) на определенном уровне. Также были разработаны теоретико-игровые методы, в которых граф представлен как некая кооперативная игра среди игроков — вершин графа, и затем находилось стабильное коалиционное разбиение. При этом предлагалось в качестве веса игроков использовать вектор Майерсона [9, 17, 18], а также теорию гедонических игр [8, 19].

Наиболее близкой к настоящей работе является статья [20], в которой было предложено применять для кластеризации графов метод максимального правдоподобия. Различие в подходах определяется тем, что в [20] анализируемый граф предполагается реализацией случайного графа с заданным распределением весов ребер, а в настоящей работе реальный граф генерируется случайным образом с заданными параметрами для внутренних и внешних связей.

В п. 2 определяется точный вид функции правдоподобия и после ряда преобразований получается простой вид целевой функции, для которой находится максимум по всевозможным разбиениям графа на подграфы. Предлагается численный алгоритм случайного поиска с использованием распределения Больцмана—Гиббса. В п. 4 приведены результаты численных расчетов для ряда примеров реальных сетей.

2. Метод максимального правдоподобия для выделения сообществ в сети. Для выделения сообществ в сети можно использовать вероятностный подход, который широко применяется в математической статистике, так называемый метод максимального правдоподобия. Следуя подходу, описанному в работе [20], запишем математическую модель выделения сообществ на основе метода максимального правдоподобия.

Предположим, что сеть генерируется случайным образом. Число сообществ фиксировано. Понятно, что теснота связей внутри сообществ должна быть более высокой, чем вне сообществ. Введем в рассмотрение следующие параметры: 1) p_{in} — вероятность возникновения связи между любыми двумя вершинами внутри сообщества; 2) p_{out} — вероятность возникновения связи между двумя вершинами из разных сообществ. Максимизируя наиболее вероятную структуру разбиения на сообщества по всем возможным конфигурациям сети, получим такое разбиение, которое соответствует реальным данным.

Рассмотрим сеть $G = (N, E)$, в которой множество вершин имеет вид $N = \{1, 2, ..., n\}$. Пусть $m = m(E)$ — число ребер в сети, $E(i, j) = 1$, если между вершинами i и j есть связь, и $E(i, j) = 0$, если нет. Назовем сообществом S непустое подмножество вершин сети, а разбиением $\Pi(N)$ — совокупность непересекающихся сообществ, объединение которых составляет в точности множество N: $\Pi(N) = \{S_1, \ldots, S_K\}$, где $\cup_{k=1}^{K} S_k = N$.

Предположим, что истинное разбиение сети $\Pi = \{S_1, ..., S_K\}$. Пусть переменные $n_k = n(S_k)$ и $m_k = m(S_k)$ обозначают число вершин и ребер в сообществе S_k, $k = 1, ..., K$, соответственно. Тогда $n = \sum_{k=1}^{K} n_k$ и $\sum_{k=1}^{K} m_k \leqslant m$.

Выразим условия, при которых разбиение на сообщества будет оптимальным.

2.1. Простой граф. Рассмотрим сообщество $S_k \in \Pi$. Вероятность реализации m_k связей среди n_k вершин в сообществе S_k равна

$$p_{in}^{m_k}(1-p_{in})^{\frac{n_k(n_k-1)}{2}-m_k}.$$

У каждой вершины i из сообщества S_k может быть в целом $n-n_k$ связей с вершинами из других сообществ, а в реальности она имеет $\sum_{j\notin S_k} E(i,j)$ связей с вершинами из других сообществ.

Вероятность реализации сети с заданной структурой равна

$$L_\Pi = \prod_{k=1}^{K} p_{in}^{m_k}(1-p_{in})^{\frac{n_k(n_k-1)}{2}-m_k} \prod_{i\in S_k} p_{out}^{\frac{1}{2}\sum_{j\notin S_k} E(i,j)} \times \\ \times (1-p_{out})^{\frac{1}{2}(n-n_k-\sum_{j\notin S_k} E(i,j))}. \tag{1}$$

Беря логарифм от функции правдоподобия L_Π (1) и упрощая его, получим

$$l_\Pi = \log L_\Pi = \sum_{k=1}^{K} m_k \log p_{in} + \sum_{k=1}^{K}\left(\frac{n_k(n_k-1)}{2} - m_k\right)\log(1-p_{in}) + \\ + \left(m - \sum_{k=1}^{K} m_k\right)\log p_{out} + \left(\frac{1}{2}\sum_{k=1}^{K} n_k(n-n_k) - \left(m - \sum_{k=1}^{K} m_k\right)\right)\log(1-p_{out}). \tag{2}$$

Разбиение Π^*, для которого функция l_Π достигает максимума по всем возможным разбиениям, назовем *оптимальным*. Заметим, что все еще остается неопределенность в выборе вероятностей p_{in} и p_{out}. Функция $l_\Pi = l_\Pi(p_{in}, p_{out})$ зависит от аргументов p_{in}, p_{out}. Максимизируя l_Π по p_{in}, p_{out}, можно затем использовать эти значения в численных расчетах.

Утверждение 1. Для фиксированного разбиения Π функция $l_\Pi(p_{in}, p_{out})$ достигает максимума при

$$p_{in} = \frac{2\sum\limits_{k=1}^{K} m_k}{\sum\limits_{k=1}^{K} n_k^2 - n}, \quad p_{out} = \frac{2\left(m - \sum\limits_{k=1}^{K} m_k\right)}{n^2 - \sum\limits_{k=1}^{K} n_k^2}. \tag{3}$$

Подставив (3) в (2), получим выражение, которое зависит только от структуры сети. Приведенные значения параметров максимизируют и функцию правдоподобия (1).

Пример 1. Рассмотрим простую сеть из восьми вершин, представленную на рис. 1.

Вычислим значение l_Π для разных разбиений. Для разбиения

$$\Pi = \{A, B, C, D, E, F, G, H\}$$

получим из (2) функцию правдоподобия

$$l_\Pi = 12\log p_{in} + 16\log(1-p_{in}).$$

Максимум этой функции достигается при $p_{in} = 3/7$. Его значение равно -19.121.

Для разбиения $\Pi = \{A, B, C, D\} \cup \{E, F, G, H\}$ имеет место функция

$$l_\Pi = 8\log p_{in} + 4\log(1-p_{in}) + 4\log p_{out} + 12\log(1-p_{out}).$$

Ее максимум достигается при $p_{in} = 2/3$ и $p_{out} = 1/4$ и равен -16.635.

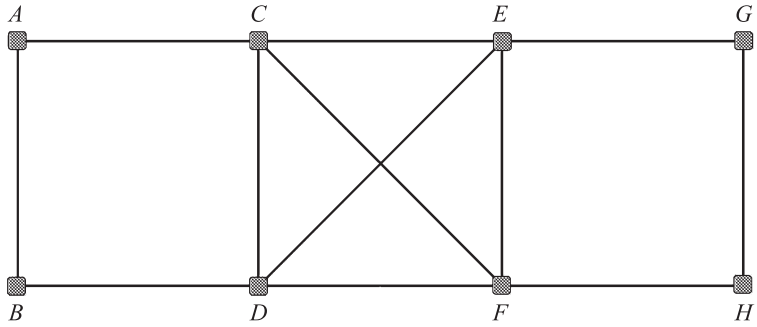

Рис. 1. Сеть из восьми вершин

Для разбиения $\Pi = \{A, B\} \cup \{C, D, E, F\} \cup \{G, H\}$ находим функцию

$$l_\Pi = 6 \log p_{in} + 4 \log p_{out} + 16 \log(1 - p_{out}).$$

Максимум этой функции достигается при $p_{in} = 1$ и $p_{out} = 1/5$. Он равен -10.008.

Видно, что разбиение $\Pi = \{A, B\} \cup \{C, D, E, F\} \cup \{G, H\}$ дает наиболее вероятную структуру сообществ для данной сети.

2.2. Мультиграф. Предположим, что между вершинами i и j возможно возникновение $m_{ij} \geqslant 0$ ребер, каждое из них с вероятностью p_{in} — если вершины находятся внутри одного сообщества и с вероятностью p_{out} — если в различных сообществах. Пусть M — наибольшее число ребер, которое может возникнуть между двумя произвольными вершинами в графе.

Таким образом, число ребер между вершинами i и j имеет мультиномиальное распределение

$$\rho(m_{ij}) = \binom{M}{m_{ij}} p^{m_{ij}} (1 - p)^{M - m_{ij}},$$

где $p = p_{in}$, если вершины i и j находятся внутри одного сообщества, и $p = p_{out}$, если в различных сообществах.

Вероятность реализации сети с заданной структурой равна

$$L_\Pi = B \prod_{k=1}^{K} p_{in}^{m_k} (1 - p_{in})^{\frac{M n_k (n_k - 1)}{2} - m_k} \prod_{i \in S_k} p_{out}^{\frac{1}{2} \sum_{j \notin S_k} m_{ij}} \times \\ \times (1 - p_{out})^{\frac{1}{2}(M(n - n_k) - \sum_{j \notin S_k} m_{ij})}, \tag{4}$$

здесь $B = \prod\limits_{\substack{i \in N \\ j \in N \\ i < j}} \binom{M}{m_{ij}}$.

Беря логарифм от функции правдоподобия L_Π (4) и упрощая его, получим функцию

$$l_\Pi = \log L_\Pi = B' + \sum_{k=1}^{K} m_k \log p_{in} + \sum_{k=1}^{K} \left(\frac{M n_k (n_k - 1)}{2} - m_k \right) \log(1 - p_{in}) + \\ + \left(m - \sum_{k=1}^{K} m_k \right) \log p_{out} + \left(\frac{1}{2} \sum_{k=1}^{K} M n_k (n - n_k) - \left(m - \sum_{k=1}^{K} m_k \right) \right) \log(1 - p_{out}),$$

в которой $B' = \log B$.

3. Связь метода максимального правдоподобия и теоретико-игровой модели. Сначала представим функцию (2) в виде

$$l_{\Pi} = \sum_{k=1}^{K} m_k \log \frac{p_{in}(1-p_{out})}{p_{out}(1-p_{in})} - \frac{1}{2}\sum_{k=1}^{K} n_k^2 \log \frac{1-p_{out}}{1-p_{in}} + R, \tag{5}$$

где выражение

$$R = -\frac{n}{2}\log(1-p_{in}) + m\log p_{out} + \left(\frac{1}{2}n^2 - m\right)\log(1-p_{out})$$

зависит от введенных параметров сети, но не зависит от конкретного разбиения. Удобно представить (5) как

$$l_{\Pi} = \log \frac{p_{in}(1-p_{out})}{p_{out}(1-p_{in})} \left(\sum_{k=1}^{K} m_k - \frac{1}{2}\sum_{k=1}^{K} n_k^2 \frac{\log \frac{1-p_{out}}{1-p_{in}}}{\log \frac{p_{in}(1-p_{out})}{p_{out}(1-p_{in})}} \right) + R.$$

Если обозначить

$$\alpha = \frac{\log \frac{1-p_{out}}{1-p_{in}}}{\log \frac{p_{in}(1-p_{out})}{p_{out}(1-p_{in})}} = \frac{\log \frac{1-p_{out}}{1-p_{in}}}{\log \frac{p_{in}}{p_{out}} + \log \frac{1-p_{out}}{1-p_{in}}}, \tag{6}$$

то метод максимального правдоподобия сведется к задаче максимизации целевой функции

$$P(\Pi) = \sum_{k=1}^{K} m_k - \frac{1}{2}\sum_{k=1}^{K} n_k^2 \alpha. \tag{7}$$

Отметим, что эта функция представляет собой потенциал в гедонической игре, связанной с данным графом, который был получен в работе [19]. Задача нахождения равновесия по Нэшу в данном случае эквивалентна задаче нахождения максимума целевой функции.

Аналогичным образом в случае графа с кратными ребрами метод максимального правдоподобия сводится к задаче максимизации целевой функции

$$P(\Pi) = \sum_{k=1}^{K} m_k - \frac{M}{2}\sum_{k=1}^{K} n_k^2 \alpha.$$

Параметр α может служить для настройки алгоритма разбиения сети. Так, в [19] рассмотрены два предельных случая $\alpha \to 0$ и $\alpha \to 1$ и доказано, что в первом из них максимум целевой функции достигается на разбиении графа, в игровой постановке соответствующем гранд-коалиции $\Pi_N = \{N\}$, а во втором — на разбиении графа на максимальные клики.

Вычислительный алгоритм поиска максимума целевой функции. Для нахождения максимума целевой функции можно применить подход, основанный на методах статистической термодинамики, а именно моделирования случайной конфигурации Π с распределением Больцмана (Гиббса)

$$\rho(\Pi) = \frac{\exp(\beta P(\Pi))}{\sum_{\forall \tilde{\Pi}} \exp(\beta P(\tilde{\Pi}))}. \tag{8}$$

При этом параметр β означает величину, обратную температуре.

Обозначим Σ множество меток сообществ в сети, $\Pi_{i\to\sigma}$ — разбиение, полученное из исходного разбиения Π переводом вершины i в сообщество σ. Алгоритм поиска максимума целевой функции заключается в следующем. Зададим число сообществ $K = K_0$ и положим некоторое разбиение Π_0 в качестве начального. Для каждой вершины $i \in N$ случайным образом выберем сообщество $\sigma \in \{1, \ldots, K\}$ и переведем ее в новое сообщество, согласно распределению вероятностей:

$$P_{\Pi\to\Pi'} = \frac{1}{n}\begin{cases} \sum\limits_{i\in N} \frac{\exp(\beta P(\Pi))}{\sum\limits_{s\in\Sigma}\exp(\beta P(\Pi_{i\to s}))}, & \text{если } \Pi' = \Pi, \\ \frac{\exp(\beta P(\Pi'))}{\sum\limits_{s\in\Sigma}\exp(\beta P(\Pi_{i\to s}))}, & \text{если } \Pi' = \Pi_{i\to\sigma}, \\ 0, & \text{иначе.} \end{cases} \tag{9}$$

Под итерацией алгоритма будем понимать n таких обновлений вершин. Известно, что случайное блуждание, описанное вероятностным распределением (9), за длительное время приведет систему в устойчивую конфигурацию [2]. При этом динамика сходимости алгоритма зависит от параметра β.

4. Численные эксперименты. Приведем результаты численных экспериментов по выделению сообществ в сети при помощи предложенного алгоритма, описанного в п. 3. Алгоритм был программно реализован в системе Wolfram Mathematica [21] (ОС Linux 64-бит). Численные эксперименты проводились на трех графах, для каждого из которых известна «истинная» структура сообществ. В каждом случае было проведено моделирование случайного блуждания по конфигурациям, имеющим распределение (8), методом Монте-Карло для нахождения приближенного решения — глобального максимума целевой функции (7), согласно алгоритму, описанному в п. 3.

Интервал изменения параметра α был выбран так, чтобы включить максимальные значения критериев качества разбиения, которых удалось достичь, а дальнейшее уменьшение шага изменения параметра не приводило бы к улучшению результатов.

4.1. Критерии оценки качества выделения сообществ. Существует множество критериев оценки качества разбиения вершин графа на сообщества. Подробный обзор таких критериев приведен в работе [22]. Для оценки эффективности описанного алгоритма рассмотрим следующие наиболее популярные критерии оценки качества разбиения:

(1) коэффициент нормированной взаимной информации (NMI);

(2) скорректированный индекс Ранда (ARI).

Данные критерии численно выражают меру близости между двумя разбиениями, поэтому применяются в том случае, когда известно «истинное» разбиение графа на сообщества.

Вычисление коэффициента NMI основано на ряде понятий из теории информации, таких как информационная энтропия и взаимная информация [23, 24]. Для разбиений Π_1 и Π_2 значение коэффициента NMI выражает количество содержащейся в них «общей» информации. Более формально его можно интерпретировать следующим образом [22]. Пусть требуется передать по коммуникационному каналу связи метки кластеров, соответствующие всем вершинам графа при разбиении Π_1. Тогда информационная энтропия $H(\Pi_1)$ интерпретируется как среднее число бит, требующееся для кодирования метки кластера каждой вершины, в соответствии с разбиением Π_1. Предположим, что получателю известно разбиение Π_2. Тогда значение коэффициента NMI выражает нормированное число бит, на которое отправитель может уменьшить $H(\Pi_1)$, сохранив при этом всю требуемую информацию.

Определим меру информационной энтропии разбиения Π, содержащего K сообществ $C_1, \ldots, C_K$, следующим образом:

$$H(\Pi) = -\sum_{k=1}^{K} \frac{|C_k|}{n} \log \frac{|C_k|}{n}$$

и меру взаимной информации между разбиениями Π_1 и Π_2, содержащими K_1 и K_2 сообществ $C_1, \ldots, C_{K_1}$ и $D_1, \ldots, D_{K_2}$:

$$MI(\Pi_1, \Pi_2) = \sum_{i=1}^{K_1} \sum_{i=1}^{K_2} \frac{|C_i \cap D_j|}{n} \log \frac{|C_i \cap D_j| n}{|C_i| \cdot |D_j|}. \tag{10}$$

Известные свойства меры взаимной информации (10) позволяют нормировать ее различными способами для удобства и в отдельных случаях повышения точности оценки. Будем использовать взаимную информацию, нормированную таким образом:

$$NMI(\Pi_1, \Pi_2) = \frac{MI(\Pi_1, \Pi_2)}{\sqrt{H(\Pi_1) H(\Pi_2)}}.$$

Коэффициент принимает значения от 0 до 1. При этом $NMI(\Pi_1, \Pi_2) = 1$ означает, что разбиения Π_1 и Π_2 идентичны друг другу, а $NMI(\Pi_1, \Pi_2) = 0$ — что разбиения совершенно различны.

Скорректированный индекс Ранда (ARI) рассчитывается на основе подсчета пар вершин, которые попадают в одинаковые или различные сообщества в двух рассматриваемых разбиениях [25, 26]. Для этого потребуются следующие величины:

- N_{11} — число пар вершин i и j, принадлежащих одному и тому же сообществу в каждом из разбиений Π_1 и Π_2;
- N_{00} — число пар вершин i и j, принадлежащих разным сообществам в каждом из разбиений Π_1 и Π_2;
- N_{10} — число пар вершин i и j, принадлежащих одному и тому же сообществу в разбиении Π_1, но разным в Π_2;
- N_{01} — число пар вершин i и j, принадлежащих одному и тому же сообществу в разбиении Π_2, но разным в Π_1.

Вычислим коэффициент ARI следующим образом:

$$ARI(\Pi_1, \Pi_2) = \frac{2\,(N_{00} N_{11} - N_{01} N_{10})}{(N_{00} + N_{01})(N_{01} + N_{11}) + (N_{00} + N_{10})(N_{10} + N_{11})}.$$

Он принимает значения от 0 до 1. При этом $ARI(\Pi_1, \Pi_2) = 1$ означает, что разбиения идентичны друг другу, а $ARI(\Pi_1, \Pi_2) = 0$ — что разбиения совершенно различны.

4.2. Сеть взаимодействий дельфинов. Рассмотрим G_d — граф коммуникаций, зафиксированных между дельфинами в географически изолированном регионе у берегов Новой Зеландии за период наблюдений с 1994 по 2001 г. [27]. Граф G_d содержит 62 вершины и 159 ребер. Вершинами графа являются дельфины; между двумя вершинами присутствует ребро, если было установлено, что данные два дельфина часто общаются. Будем предполагать, что дельфины внутри одного и того же сообщества в искомом разбиении Π^* общаются друг с другом чаще, чем дельфины из разных сообществ.

Структура сообществ, которую будем считать «истинной» (рис. 2), была вычислена в работе [16] на основе центральности по кратчайшему пути. Обозначим соответствующее разбиение вершин Π_{gt}. Оно отвечает в реальности наблюдаемому

биологами разделению дельфинов на две группы, связанные небольшим числом контактов. Более крупная из двух групп подразделяется еще на три группы: одна из них состоит преимущественно из самок, а две — из самцов.

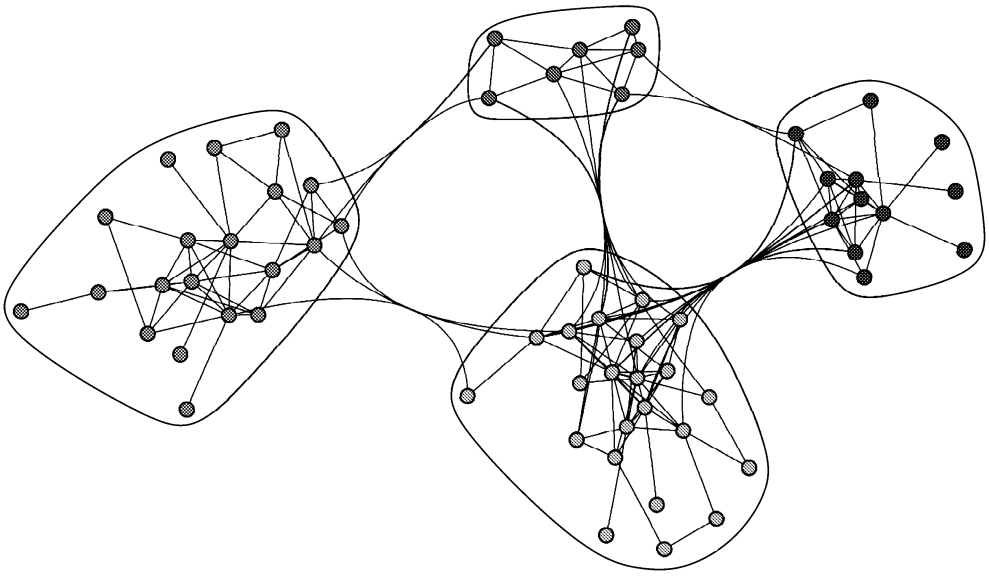

Рис. 2. Структура сообществ в социальной сети дельфинов [16]

Было произведено моделирование случайного блуждания по конфигурациям методом Монте-Карло для нахождения приближенного решения — глобального максимума целевой функции (7). В качестве начального было задано разбиение вершин графа на 4 равных по величине сообщества.

Таблица 1. **Поиск глобального максимума целевой функции на графе социальных взаимодействий дельфинов,** $\alpha \in [0.0, \ldots, 0.15]$

α	K	$P(\alpha, \Pi_{gt})$	$P(\alpha, \Pi^*)$	$NMI(\Pi_{gt}, \Pi^*)$	$ARI(\Pi_{gt}, \Pi^*)$
0.00	1	129.0	159.0	0.000	0.000
0.01	3	123.4	142.4	0.630	0.457
0.02	3	117.8	131.3	0.595	0.437
0.03	4	112.2	121.7	0.563	0.452
0.04	4	106.6	108.6	0.793	0.719
0.05	4	101.0	101.7	0.853	0.775
0.06	4	95.3	96.0	0.726	0.649
0.07	4	89.7	91.0	0.831	0.812
0.08	4	84.1	85.8	0.831	0.812
0.09	4	78.5	179.3	0.749	0.703
0.10	4	72.9	75.5	0.788	0.774
0.11	4	67.3	70.3	0.762	0.740
0.12	4	61.7	64.4	0.810	0.809
0.13	4	56.1	60.4	0.762	0.740
0.14	4	50.5	55.3	0.715	0.694
0.15	4	44.9	50.4	0.741	0.707

В табл. 1 приведены результаты численных экспериментов по поиску разбиения Π^*, максимизирующего целевую функцию (7). В каждом эксперименте было задано

значение α, параметр β принят равным 2.5, 5, 10 и 15 — по 250 итераций на каждое значение. Решение для $\alpha = 0$ было найдено аналитически.

Для каждого значения параметра α в табл. 1 указаны: K — полученное число сообществ, $P(\alpha, \Pi_{gt})$ — значение целевой функции на «истинном» разбиении, $P(\alpha, \Pi^*)$ — найденный максимум целевой функции, $NMI(\Pi_{gt}, \Pi^*)$ и $ARI(\Pi_{gt}, \Pi^*)$ — оценки качества найденного разбиения, описанные в п. 4.1.

4.3. Сеть футбольных матчей. Рассмотрим G_f — граф матчей по американскому футболу, проведенных в осенний сезон 2000 г. [11]. Граф G_f содержит 115 вершин и 613 ребер. Вершинами графа являются команды; между двумя вершинами присутствует ребро, если состоялась игра между данными командами. Будем предполагать, что команды внутри одного и того же сообщества в искомом разбиении Π^* играют друг с другом чаще, чем команды из разных сообществ.

Структура сообществ Π_{gt}, которую будем считать «истинной» (рис. 3), состоит в распределении команд по 12 конференциям.

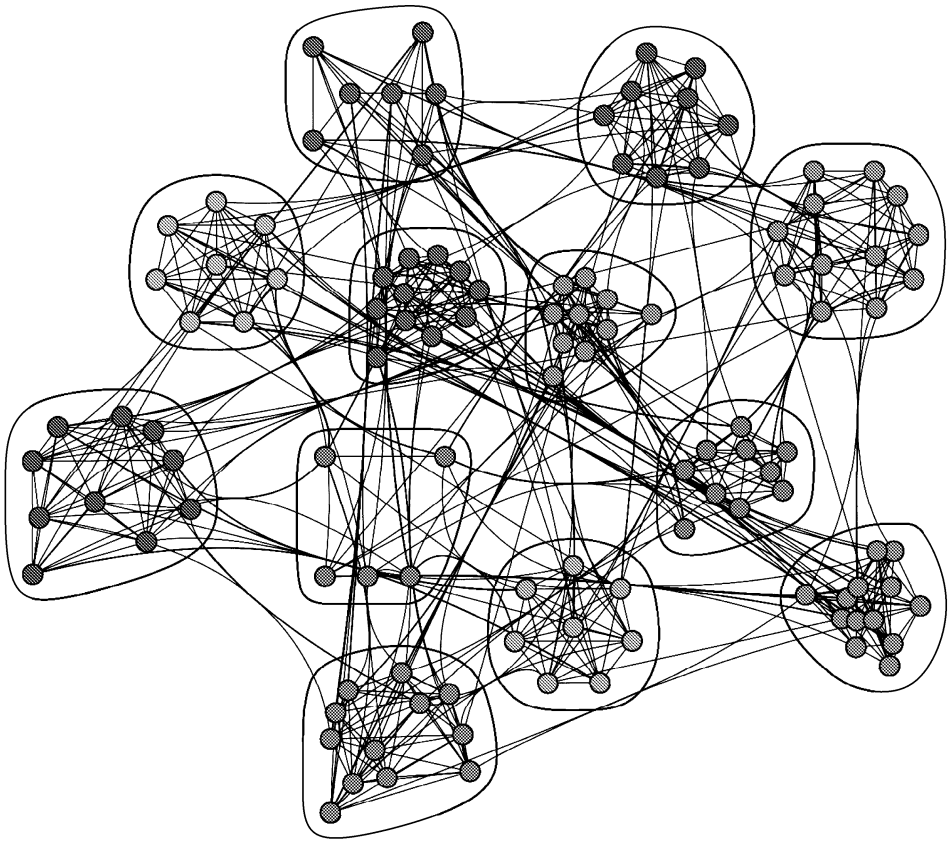

Рис. 3. Структура сообществ в сети футбольных матчей [11]

Было проведено моделирование случайного блуждания по конфигурациям методом Монте-Карло для нахождения приближенного решения — глобального максимума целевой функции (7). В качестве начального было задано разбиение вершин графа на 12 равных по величине сообществ. В табл. 2 приведены результаты численных экспериментов по поиску разбиения Π^*, максимизирующего целевую функцию (7).

В каждом эксперименте было задано значение α, параметр β принят равным 2.5, 5, 10 и 15 — по 250 итераций на каждое значение. Решение для $\alpha = 0$ было найдено аналитически.

Таблица 2. **Поиск глобального максимума целевой функции на сети футбольных матчей,** $\alpha \in [0.0, \ldots, 0.05]$

α	K	$P(\alpha, \Pi_{gt})$	$P(\alpha, \Pi^*)$	$NMI(\Pi_{gt}, \Pi^*)$	$ARI(\Pi_{gt}, \Pi^*)$
0.0	1	394	613	0	0
0.1	10	336	362.6	0.889	0.814
0.2	10	277.9	293.3	0.902	0.847
0.3	12	219.9	247.1	0.924	0.906
0.4	12	161.8	182.2	0.824	0.872
0.5	12	103.8	130.8	0.931	0.915
0.6	12	45.7	72.9	0.931	0.915
0.7	12	−12.4	15.1	0.931	0.915
0.8	12	−70.4	−44	0.919	0.870
0.9	12	−128.5	−100.1	0.909	0.860
1.0	12	−186.5	−156.5	0.909	0.860

4.4. Сеть городского транспорта. Рассмотрим G_t — транспортную сеть города Петрозаводск. Вершинами графа G_t являются остановочные пункты общественного транспорта; между двумя вершинами присутствует ребро, если два остановочных пункта напрямую связаны автодорогой. Граф G_t содержит 136 вершин и 204 ребра. Согласно математической модели, можно предполагать, что остановочные пункты связаны большим числом дорог внутри одного и того же района города; напротив, остановочные пункты разных районов связаны меньшим числом дорог, поскольку границы районов проходят по рекам, железным дорогам, скоростным шоссе и т. п., а число переходов через них (таких как мосты) ограничено.

В соответствии с административно-территориальным делением город состоит из 16 микрорайонов. Будем считать соответствующее разбиение вершин на сообщества «истинным» (рис. 4).

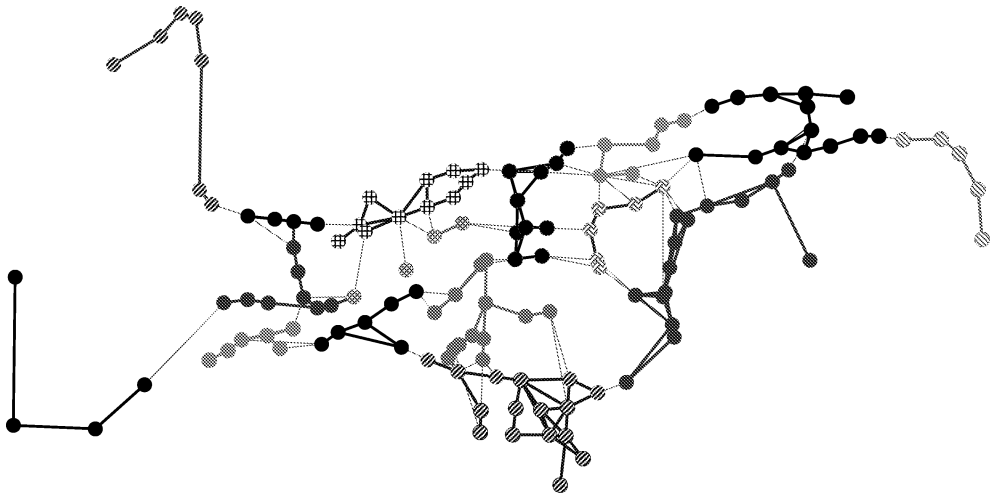

Рис. 4. Структура сообществ в транспортной сети Петрозаводска

Для того чтобы учесть расстояние между остановочными пунктами, будем рассматривать G_t как мультиграф, полагая число ребер между вершинами i и j, напрямую связанных автодорогой, равным

$$m_{ij} = \frac{d_{\max}}{d_{ij}},$$

где d_{ij} — длина участка автодороги между остановочными пунктами i и j; $d_{\max}$ — максимальная длина таких участков в транспортной сети.

Как и в п. 4.3, было произведено моделирование случайного блуждания по конфигурациям методом Монте-Карло для нахождения приближенного решения — глобального максимума целевой функции (7).

В табл. 3 приведены результаты численных экспериментов по поиску разбиения Π^*, максимизирующего целевую функцию (7). В качестве начального было задано разбиение вершин графа на 16 равных по величине сообществ. В каждом эксперименте было задано значение α, параметр β принят равным 2.5, 5, 10 и 15 — по 250 итераций на каждое значение. Решение для $\alpha = 0$ было найдено аналитически.

Таблица 3. **Поиск глобального максимума целевой функции на транспортной сети,** $\alpha \in [0.0, \ldots, 0.05]$

α	K	$P(\alpha, \Pi_{gt})$	$P(\alpha, \Pi^*)$	$NMI(\Pi_{gt}, \Pi^*)$	$ARI(\Pi_{gt}, \Pi^*)$
0.000	1	780.3	1013.4	0	0
0.010	16	567.5	659.1	0.739	0.507
0.015	16	461.1	549.7	0.723	0.493
0.020	16	354.7	480	0.765	0.549
0.025	16	248.4	365.3	0.719	0.460
0.030	16	142	288	0.700	0.447
0.035	16	35.6	190.6	0.717	0.466
0.040	16	−70.79	84.9	0.686	0.408
0.045	16	−177.2	5.9	0.705	0.454
0.050	16	−283.6	−73.4	0.692	0.434

На рис. 5 представлена динамика поиска глобального максимума целевой функции для трех рассмотренных графов. В каждом эксперименте параметр β был равен 2.5, 5, 10 и 15 — по 250 итераций на каждое значение. Во всех экспериментах глобальный максимум был найден в среднем за 500–600 итераций алгоритма.

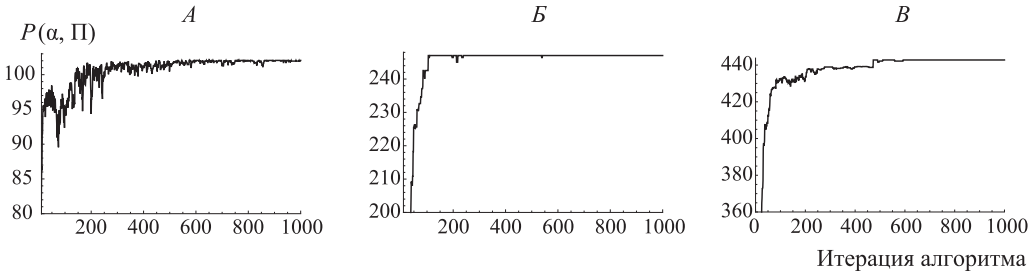

Рис. 5. Динамика поиска глобального максимума целевой функции
A — G_d, $n = 62$; $Б$ — G_f, $n = 115$; B — G_t, $n = 136$.

Результаты экспериментов показали, что в случае каждого из изученных графов и для всех значений параметра α «истинное» разбиение не максимизирует целевую функцию. Однако найденное разбиение оказывается близким к «истинному»

по рассмотренным критериям. Таким образом, разработанный алгоритм позволяет получить приближенное решение. Далее оценим качество этого решения в сравнении с двумя известными алгоритмами для выделения структуры сообществ.

В табл. 4 приведены результаты численных экспериментов по поиску структуры сообществ в графе. Для каждого из данных графов сравниваются три разбиения: найденное при помощи предложенного алгоритма как приближение глобального максимума целевой функции, разбиение на основе модулярности и разбиение на основе центральности [12].

Таблица 4. **Результаты выделения сообществ в графе методом максимального правдоподобия**

Граф	Разбиение Π	K	$NMI(\Pi_{gt}, \Pi)$	$ARI(\Pi_{gt}, \Pi)$	Время счета, с
G_d		4	0.853	0.775	11
	Разбиение на основе модулярности	4	0.703	0.668	<1
	Разбиение на основе центральности	5	0.912	0.894	<1
G_f		12	0.931	0.915	24
	Разбиение на основе модулярности	7	0.770	0.564	<1
	Разбиение на основе центральности	10	0.880	0.796	<1
G_t		16	0.748	0.554	72
	Разбиение на основе модулярности	12	0.712	0.501	<1
	Разбиение на основе центральности	12	0.725	0.512	<1

Второй и третий алгоритмы реализованы в ядре системы Wolfram Mathematica, поэтому скорость их работы существенно меньше, чем скорость предложенного алгоритма. Тем не менее результаты, представленные в табл. 4, показывают, что для графов G_f и G_t такой алгоритм позволил найти разбиение, наиболее близкое к «истинному» по критериям *NMI* и *ARI*. Для графа G_d построенное разбиение оказалось более близким к «истинному», чем разбиение на основе модулярности, но все же менее близким, чем разбиение на основе центральности. Отметим, что данный результат не свидетельствует о неприменимости разработанного алгоритма и объясняется структурой «истинного» разбиения — оно построено именно на основе одного из видов центральности вершин.

5. Заключение. Графы, представляющие реальные социальные и коммуникационные сети, быстро изменяются, при этом эффективным инструментом их изучения служат случайные графы. Важной задачей является выделить структуру сообществ в сетях. В условиях большой размерности сетей особенно актуальны приближенные методы, которые позволяют за ограниченное время находить решение, близкое к оптимальному.

В данной статье описана математическая модель, в которой граф генерируется случайным образом с заданными параметрами для внутренних и внешних связей между вершинами, а сообщества полагаются непересекающимися. Предложен метод выделения структуры сообществ на основе метода максимального правдоподобия, и на его основе описан численный алгоритм случайного поиска с использованием распределения Больцмана—Гиббса. Приведены и анализируются результаты численных расчетов для трех примеров реальных сетей.

Расчеты для реальных сетей малой размерности позволили прийти к выводам об эффективности предложенного алгоритма при выборе подходящих значений его параметров, зависящих от структуры сети. Параметр α задан аналитическим выражением (6), которое зависит от механизмов, лежащих в основе формирования сети.

Параметр β играет роль обратной температуры в ходе случайного поиска и определяет амплитуду колебаний целевой функции. Общее число итераций алгоритма также является параметром, который зависит от желаемого баланса между скоростью и точностью поиска. При анализе сетей большой размерности возможна адаптивная настройка значений параметров алгоритма на основе информации, полученной на предыдущих шагах. При этом начальные оценки параметров могут быть получены на основе априорных представлений о структуре конкретной сети.

В настоящее время наибольший интерес представляет разработка математической модели, описывающей структуру пересекающихся сообществ в графе. Такая структура сообществ свойственна для современных сложных сетей большой размерности. Кроме того, эти сети зачастую развиваются динамически (например, представляя собой сеть движущихся объектов, снабженных датчиками), что задает еще одно направление дальнейших исследований. В случае динамической сети математическая модель должна позволить выделить ключевые группы вершин, динамика которых задает изменения остальной части сети.

Литература

1. *Freeman L. C.* A set of measures of centrality based on betweenness // Sociometry. 1977. Vol. 40. P. 35–41.

2. *Levin D. A., Peres Y.* Markov chains and mixing times. Providence, Rhode Island: Amer. Mathematical Soc., 2017. 447 p.

3. *Page L., Brin S., Motwani R., Winograd T.* The PageRank citation ranking: Bringing order to the web.: technical report. Stanford: Stanford InfoLab, 1998. 17 p.

4. *Pons P., Latapy M.* Computing communities in large networks using random walks // Journal of Graph Algorithms and Applications. 2006. Vol. 10(2). P. 191–218.

5. *Avrachenkov K. E., Mazalov V. V., Tsynguev B. T.* Beta current flow centrality for weighted networks // Proceedings of CSoNET–2015, LNCS. 2015. Vol. 9197. P. 216–227.

6. *Brandes U., Fleischer D.* Centrality measures based on current flow // Proceedings of the 22nd annual conference on Theoretical Aspects of Computer Science. 2005. P. 533–544.

7. *Mazalov V., Tsynguev B.* Kirchhoff centrality measure for collaboration network // CSoNet–2016. LNCS, 2016. Vol. 9795. P. 147–157.

8. *Bogomolnaia A., Jackson M. O.* The stability of hedonic coalition structures // Games and Economic Behavior. 2002. Vol. 38(2). P. 201–230.

9. *Mazalov V. V., Avrachenkov K. E., Trukhina L. I., Tsynguev B. T.* Game-theoretic centrality measures for weighted graphs // Fundamenta Informaticae. 2016. Vol. 145(3). P. 341–358.

10. *Fortunato S., Barthelemy M.* Resolution limit in community detection // Proceedings of the National Academy of Sciences USA. 2007. Vol. 104(1). P. 36–41.

11. *Girvan M., Newman M. E. J.* Community structure in social and biological networks // Proceedings of the National Academy of Sciences USA. 2002. Vol. 99(12). P. 7821–7826.

12. *Fortunato S.* Community detection in graphs // Physics Reports. 2010. Vol. 486(3). P. 75–174.

13. *Zachary W. W.* An information flow model for conflict and fission in small groups // Journal of Anthropological Research. 1977. Vol. 33(4). P. 452–473.

14. *Jackson M. O.* Social and economic networks. Princeton: Princeton University Press, 2010. 520 p.

15. *Kaur R., Singh S.* A survey of data mining and social network analysis based anomaly detection techniques // Egypt Inf. Journal. 2016. Vol. 17(2). P. 199–216.

16. *Newman M. E., Girvan M.* Finding and evaluating community structure in networks // Physical Review E. 2004. Vol. 69(2). P. 026113.

17. *Meila M., Shi J.* A Random walks view of spectral segmentation // Proceedings of AISTATS. 2001. P. 1–6.

18. *Myerson R. B.* Graphs and cooperation in games // Math. Oper. Res. 1977. Vol. 2. P. 225–229.

19. *Avrachenkov K., Kondratev A., Mazalov V.* Cooperative Game Theory approaches for network partitioning // Computing and Combinatorics / eds.: Y. Cao, J. Chen. COCOON. 2017. LNCS. 2017. Vol. 10392. P. 591–603.

20. *Copic J., Jackson M., Kirman A.* Identifying community structures from network data via maximum likelihood methods // The B. E. Journal of Theoretical Economics. 2009. Vol. 9, iss. 1. P. 1635–1704.

21. Wolfram Research, Inc. (www.wolfram.com). Mathematica Online. Champaign, IL., 2018.

22. *Vinh N. X., Epps J., Bailey J.* Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance // Journal of Machine Learning Research. 2010. Vol. 11. P. 2837–2854.

23. *Kvalseth T. O.* Entropy and correlation: Some comments // Systems, Man and Cybernetics, IEEE Transactions on. 1987. Vol. 17(3). P. 517–519.

24. *Strehl A., Ghosh J.* Cluster ensembles – a knowledge reuse framework for combining multiple partitions // Journal of Machine Learning Research. 2002. Vol. 3. P. 583–617.

25. *Hubert L., Arabie P.* Comparing partitions // Journal of Classification. 1985. Vol. 2(1). P. 193–218.

26. *Steinley D.* Properties of the Hubert-Arabie adjusted Rand index // Psychol. Methods. 2004. Vol. 9(3). P. 386–396.

27. *Lusseau D., Schneider K., Boisseau O. J.* et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations // Behav. Ecol. Sociobiol. 2003. Vol. 54. P. 396–405.

Статья поступила в редакцию 30 мая 2018 г.; принята к печати 14 июня 2018 г.

Контактная информация:

Мазалов Владимир Викторович — д-р физ.-мат. наук, проф.; vmazalov@krc.karelia.ru

Никитина Наталия Николаевна — канд. техн. наук; nikitina@krc.karelia.ru

The maximum likelihood method for detecting communities in communication networks

V. V. Mazalov[1,2], *N. N. Nikitina*[2]

[1] St. Petersburg State University, 7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

[2] Federal Research Center "Karelian Research Center of the Russian Academy of Sciences", 11, Pushkinskaya ul., Petrozavodsk, 185910, Russian Federation

For citation: Mazalov V. V., Nikitina N. N. The maximum likelihood method for detecting communities in communication networks. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2018, vol. 14, iss. 3, pp. 200–214. https://doi.org/10.21638/11702/spbu10.2018.302

The community detection in social and communication networks is an important problem in many applied fields: biology, sociology, social networks. This is especially true for networks that are represented by large graphs. In this paper, we propose a method for community detection based on the maximum likelihood method for the random formation of a network with given parameters of the tightness of connections within the community and between different communities. A numerical algorithm for finding the maximum of the objective function over all possible network partitions is described. The algorithm is implemented and tested on real networks of small dimension.

Keywords: network communities, detecting communities in a network, maximum likelihood method, Gibbs sampling.

References

1. Freeman L. C. A set of measures of centrality based on betweenness. *Sociometry*, 1977, vol. 40, pp. 35–41.

2. Levin D. A., Peres Y. *Markov chains and mixing times*. Providence, Rhode Island, Amer. Mathematical Soc. Publ., 2017, 447 p.

3. Page L., Brin S., Motwani R., Winograd T. *The PageRank citation ranking: Bringing order to the web*. Technical Report. Stanford, Stanford InfoLab Publ., 1998, 17 p.

4. Pons P., Latapy M. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 2006, vol. 10(2), pp. 191–218.

5. Avrachenkov K. E., Mazalov V. V., Tsynguev B. T. Beta current flow centrality for weighted networks. *Proceedings of CSoNET–2015*, LNCS, 2015, vol. 9197, pp. 216–227.

6. Brandes U., Fleischer D. Centrality measures based on current flow. *Proceedings of the 22nd annual conference on Theoretical Aspects of Computer Science*, 2005, pp. 533–544.

7. Mazalov V., Tsynguev B. Kirchhoff centrality measure for collaboration network. *CSoNet–2016*, LNCS, 2016, vol. 9795, pp. 147–157.

8. Bogomolnaia A., Jackson M. O. The stability of hedonic coalition structures. *Games and Economic Behavior*, 2002, vol. 38(2), pp. 201–230.

9. Mazalov V. V., Avrachenkov K. E., Trukhina L. I., Tsynguev B. T. Game-theoretic centrality measures for weighted graphs. *Fundamenta Informaticae*, 2016, vol. 145(3), pp. 341–358.

10. Fortunato S., Barthelemy M. Resolution limit in community detection. *Proceedings of the National Academy of Sciences USA*, 2007, vol. 104(1), pp. 36–41.

11. Girvan M., Newman M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences USA*, 2002, vol. 99(12), pp. 7821–7826.

12. Fortunato S. Community detection in graphs. *Physics Reports*, 2010, vol. 486(3), pp. 75–174.

13. Zachary W. W. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 1977, vol. 33(4), pp. 452–473.

14. Jackson M. O. *Social and economic networks*. Princeton, Princeton University Press, 2010, 520 p.

15. Kaur R., Singh S. A survey of data mining and social network analysis based anomaly detection techniques. *Egypt Inf. Journal*, 2016, vol. 17(2), pp. 199–216.

16. Newman M. E., Girvan M. Finding and evaluating community structure in networks. *Physical Review E*, 2004, vol. 69(2), pp. 026113.

17. Meila M., Shi J. A Random Walks View of spectral segmentation. *Proceedings of AISTATS*, 2001, pp. 1–6.

18. Myerson R. B. Graphs and cooperation in games. *Math. Oper. Res.*, 1977, vol. 2, pp. 225–229.

19. Avrachenkov K., Kondratev A., Mazalov V. Cooperative Game Theory approaches for network partitioning. *Computing and Combinatorics*. Eds. by Y. Cao, J. Chen. COCOON, 2017, LNCS, 2017, vol. 10392, pp. 591–603.

20. Copic J., Jackson M., Kirman A. Identifying community structures from network data via maximum likelihood methods. *The B. E. Journal of Theoretical Economics*, 2009, vol. 9, iss. 1, pp. 1935–1704.

21. *Wolfram Research, Inc.* (www.wolfram.com). Mathematica Online. Champaign, IL., 2018.

22. Vinh N. X., Epps J., Bailey J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 2010, vol. 11, pp. 2837–2854.

23. Kvalseth T. O. Entropy and correlation: Some comments. *Systems, Man and Cybernetics, IEEE Transactions on*, 1987, vol. 17(3), pp. 517–519.

24. Strehl A., Ghosh J. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 2002, vol. 3, pp. 583–617.

25. Hubert L., Arabie P. Comparing partitions. *Journal of Classification*, 1985, vol. 2(1), pp. 193–218.

26. Steinley D. Properties of the Hubert-Arabie adjusted Rand index. *Psychol. Methods*, 2004, vol. 9(3), pp. 386–396.

27. Lusseau D., Schneider K., Boisseau O. J. et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behav. Ecol. Sociobiol.*, 2003, vol. 54, pp. 396–405.

Author's information:

Vladimir V. Mazalov — Dr. Sci. in Physics and Mathematics, Professor; vmazalov@krc.karelia.ru

Natalia N. Nikitina — PhD Sci. in Technics; nikitina@krc.karelia.ru