УДК 004.421.2:519.178
DOI 10.25205/1818-7900-2017-15-3-64-73

А. О. Орлов, **А. А. Чеповский**

Национальный исследовательский университет – Высшая школа экономики
ул. Мясницкая, 20, Москва,101000, Россия

aachepovsky@hse.ru

О СВОЙСТВАХ МОДУЛЯРНОСТИ И АКТУАЛЬНЫХ КОРРЕКТИРОВКАХ АЛГОРИТМА БЛОНДЕЛЯ [*]

Одной из задач, связанных с изучением сложных сетей, является задача выявления структуры сообществ – разбиения всех вершин на группы (сообщества), таким образом, чтобы вершины каждой группы были более плотно связаны между собой, нежели с остальным графом. Популярным алгоритмом выделения сообществ является алгоритм Блонделя, основанный на максимизации модулярности Ньюмана – Гирван, распространенного критерия оценки качества разбиений на сообщества. Данная статья посвящена анализу его особенностей и результатов работы, а также возможным модификациям. Разобраны результаты тестирования как на сгенерированных графах, так и на реальных данных.

Ключевые слова: структура графа, анализ социальной сети, выделение сообществ, большие данные.

Граф сети взаимодействующих объектов

В данной работе под сетью взаимодействующих объектов далее понимается граф пользователей, полученный через открытый API социальных сетей. Вершины данного графа соответствуют аккаунтам пользователей, а ребра в зависимости от социальной сети отвечают либо отношениям «дружба», тогда получается неориентированный граф, либо отношениям «подписка», тогда получается ориентированный граф. Далее рассматриваются особенности разбиения множества вершин на основании топологии данного графа на неявные сообщества.

Модулярность

При рассмотрении графа, имеющего структуру сообществ, ожидается, что плотность связей внутри этих сообществ будет выше, чем плотность связей в остальном графе. Также можно ожидать, что плотность связей внутри сообществ в таком графе должна быть выше, чем плотность связей в графе, не обладающем структурой сообществ. Случайный граф, моделирующий некоторые свойства исходной сети, но не обладающий структурой сообществ, называется нулевой моделью. Сравнивая внутрикластерную плотность оригинального графа с ожидаемой в нулевой модели, можно ввести функционал качества разбиений, называемый модулярностью:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - P_{ij}\right)\cdot\delta\left(C_i, C_j\right),$$

[*] Работа выполнена при поддержке РФФИ, гранты № 16-29-09546 и 16-07-00641.

Орлов А. О., *Чеповский А. А.* О свойствах модулярности и актуальных корректировках алгоритма Блонделя // Вестн. НГУ. Серия: Информационные технологии. 2017. Т. 15, № 3. С. 64–73.

где A – матрица смежности графа; P_{ij} – ожидаемое число ребер между вершинами i и j в графе, не обладающем структурой сообществ; m – число ребер в графе; C_i – сообщество, к которому принадлежит i-я вершина; δ – дельта-функция.

Таким образом, значение модулярности напрямую зависит от выбора нулевой модели.

Модулярность Ньюмана – Гирван

Наиболее популярной нулевой моделью является модель Ньюмана – Гирван [1–3]. Число ребер P_{ij} в данной модели считается по следующей формуле:

$$P_{ij} = \frac{d_i \cdot d_j}{2m},$$

где d_i – степень i-й вершины; m – число ребер в графе.

Данная модель сохраняет степени вершин графа и при этом предполагает случайное распределение ребер между ними. В этой модели модулярность определяется по следующей формуле:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{d_i \cdot d_j}{2m}\right)\cdot \delta\left(C_i, C_j\right).$$

Отметим, что данная формула легко обобщается на случай взвешенных графов: в таком случае d_i определяется как сумма весов ребер, инцидентных i-й вершине (петля учитывается дважды), а m – как сумма весов всех ребер.

Алгоритмы выделения сообществ

Модулярность Ньюмана – Гирван является одной из наиболее популярных метрик качества для разбиений. На ее максимизации основаны многие алгоритмы выделения сообществ. Однако следует отметить, что поиск глобального максимума модулярности – это *NP*-полная задача в сильном смысле [4]. Однако существуют эвристические алгоритмы, хорошо оптимизирующие модулярность. Одним из них является алгоритм Блонделя [5], эффективно оптимизирующий функционал модулярности и обладающий низкой временной сложностью: $O\left(n \cdot \log\left(n\right)\right)$, где n – число вершин в графе.

Алгоритм Блонделя

Ниже представлено описание алгоритма Блонделя.

1. Присвоить каждой вершине i собственное сообщество $C_i \to i$.
2. Для каждой вершины i:
   а) для каждого ее соседа j подсчитать изменение модулярности при перемещении вершины i из своего сообщества в сообщество C_j и найти максимальный положительный прирост $\max_j \Delta Q$;
   б) вершина i переносится в сообщество, дающее максимальный прирост модулярности.
3. Повторять шаг 2 до тех пор, пока Q увеличивается.
4. В случае, если значение модулярности на втором шаге изменялось:
   а) создать метаграф путем объединения вершин каждого сообщества в метавершины; вес ребра между двумя метавершинами будет равен весу ребер между соответствующими сообществами в исходном графе, а сумма весов ребер внутри сообщества будет представлена в виде петли с тем же весом;
   б) запустить алгоритм на полученном метаграфе.

Как видно из описания алгоритма, для каждой вершины необходимо подсчитать изменение модулярности при переносе ее из своего сообщества в сообщество каждой из соседних вершин. Изменение модулярности в таком случае складывается из двух компонент: изменение от удаления вершины i из ее собственного сообщества C_i (в таком случае вершина связывается с новым сообществом C_k, содержащим только эту вершину) и изменение от добавления вершины в новое сообщество C_j. Изменение модулярности при добавлении вершины i в сообщество C рассчитывается по следующей формуле:

$$\Delta Q = \frac{k_i^C}{m} - \frac{\Sigma_{\text{tot}}^C \cdot d_i}{2m^2},$$

где k_i^C – сумма весов ребер, инцидентных вершине i и сообществу C; m – сумма весов ребер графа; Σ_{tot}^C – сумма степеней вершин, принадлежащих сообществу C; d_i – степень вершины i.

Алгоритм Блонделя для ориентированного графа

Как уже было сказано, алгоритм Блонделя основывается на оптимизации модулярности Ньюмана – Гирван, которая определяется для неориентированного графа. Одним из возможных вариантов работы с ориентированным графом является его преобразование в неориентированный. Но при таких модификациях теряется часть информации о сети. Иным решением является использование ориентированной модулярности при работе алгоритма [6], которая определяется следующим образом:

$$Q_d = \frac{1}{m}\sum_{ij}\left(A_{ij} - \frac{d_i^{\text{in}} \cdot d_j^{\text{out}}}{m}\right)\delta\left(C_i, C_j\right),$$

где A – матрица смежности графа; d_i^{in} – полустепень захода *i*-й вершины; d_j^{out} – полустепень исхода *j*-й вершины. Тогда расчет ΔQ_d при добавлении *i*-й вершины в сообщество C будет производиться по формуле

$$\Delta Q_d = \frac{k_i^C}{m} - \frac{d_i^{\text{in}} \cdot \Sigma_{\text{out}}^C + d_i^{\text{out}} \cdot \Sigma_{\text{in}}^C}{m^2},$$

где Σ_{out}^C – сумма полустепеней исхода всех вершин, входящих в сообщество C; Σ_{in}^C – сумма полустепеней захода всех вершин, входящих в сообщество C.

Особенности модулярности и ее подсчета в алгоритме Блонделя

Так как алгоритм Блонделя оптимизирует функционал модулярности, особенности результатов применения этого алгоритма связаны с некоторыми свойствами модулярности. Рассмотрим в неориентированном случае выражение для изменения модулярности при переносе вершины *i* из тривиального сообщества (сообщества, состоящего лишь из этой вершины) в сообщество *C*:

$$\Delta Q = \frac{k_i^C}{m} - \frac{\Sigma_{\text{tot}}^C \cdot d_i}{2m^2}.$$

Получаем следующий критерий для переноса вершины из тривиального сообщества в сообщество *C*:

$$k_i^C - \frac{\Sigma_{\text{tot}}^C \cdot d_i}{2m} > 0.$$

Покажем, что разбиение, максимизирующее функционал модулярности на связном простом графе, не содержит тривиальных сообществ.

Пусть $G(V,E)$ – простой связный граф на n $(n>1)$ вершинах и пусть $C=\{c_{v_1},...,c_{v_n}\}$ – некоторое разбиение этого графа на сообщества (c_{v_i} – номер сообщества вершины v_i). Пусть также существует сообщество $c_{v_k}\in C,$ содержащее только вершину v_k. Покажем, что в данном разбиении существует вершина $v_i,$ инцидентная $v_k,$ такая что при переносе v_k в сообщество v_i модулярность разбиения увеличится. Тогда мы докажем отсутствие тривиальных сообществ в максимальном по модулярности разбиении.

Введем вспомогательное обозначение: назовем сумму степеней вершин в *i*-м сообществе степенью *i*-го сообщества, и обозначим ее как Σ_i^C. Также обозначим через *d* степень вершины v_k. По условию связности графа $d>0.$ Таким образом, вершина v_k имеет *d* соседей и связана максимум с *d* сообществами, причем ровно с *d* только в том случае, если все вершины-соседи лежат в различных сообществах.

Рассмотрим этот случай отдельно. Обозначим за Σ_m^C минимальную степень сообщества среди всех сообществ-соседей вершины $v_k,$ а за *m* – номер сообщества с такой степенью. Запишем критерий переноса вершины v_k в сообщество *m*:

$$k_i^C-\frac{\Sigma_m^C\cdot d}{2m}>0.$$

Но в данном случае $k_i^C=1,$ также в силу минимальности степени у данного сообщества для общего числа связей имеем $2m\ge d+\Sigma_m^C\cdot d=\Sigma_m^C\cdot(d+1),$ откуда получаем

$$1-\frac{\Sigma_m^C\cdot d}{2m}\ge 1-\frac{\Sigma_m^C\cdot d}{\Sigma_m^C\cdot(d+1)}=1-\frac{d}{d+1}>0.$$

Теперь докажем исходное утверждение для любого количества сообществ-соседей. Пусть вершина v_k имеет в рассматриваемом разбиении *q* сообществ-соседей $(q\le d).$ В таком случае вершина v_k может быть связана с некоторыми сообществами сразу несколькими ребрами. Тогда мы можем разделить все сообщества-соседи на группы по числу их связей с вершиной. Обозначим количество таких групп через $t>0,$ число сообществ в *i*-й группе – $n_i>0,$ число связей – $l_i>0,$ минимальную степень сообщества среди всех сообществ в группе – $w_{m_i}>0,$ где номер самого сообщества из *i*-й группы обозначен за m_i. Получаем

$$\sum_{i=1}^{t}n_i=q,$$

$$\sum\nolimits_{i=1}^{t}(n_i\cdot l_i)=d,$$

$$2m\ge d+\sum_{i=1}^{t}(w_{m_i}\cdot n_i).$$

Тогда для переноса вершины в сообщество m_i должно быть верным неравенство

$$l_i-\frac{w_{m_i}\cdot d}{d+\sum\nolimits_{i=1}^{t}(w_{m_i}\cdot n_i)}>0.$$

Покажем, что всегда существует *i*, для которого это неравенство выполняется. Предположим, что такого *i* не существует, т. е. для всех *i* верно

$$l_i-\frac{w_{m_i}\cdot d}{d+\sum\nolimits_{i=1}^{t}(w_{m_i}\cdot n_i)}\le 0.$$

Тогда, домножив на n_i каждое неравенство, получим $(n_i>0)$:

$$l_i\cdot n_i-\frac{n_i\cdot w_{m_i}\cdot d}{d+\sum\nolimits_{i=1}^{t}(w_{m_i}\cdot n_i)}\le 0.$$

Теперь сложим все полученные неравенства:

$$\sum_{i=1}^{t}\left(n_i \cdot l_i\right) - \frac{\sum_{i=1}^{t}\left(n_i \cdot w_{m_i} \cdot d\right)}{d + \sum_{i=1}^{t}\left(w_{m_i} \cdot n_i\right)} \le 0.$$

Упростим последнее выражение:

$$d - d\frac{\sum_{i=1}^{t}\left(w_{m_i} \cdot n_i\right)}{d + \sum_{i=1}^{t}\left(w_{m_i} \cdot n_i\right)} = d\left(1 - \frac{\sum_{i=1}^{t}\left(w_{m_i} \cdot n_i\right)}{d + \sum_{i=1}^{t}\left(w_{m_i} \cdot n_i\right)}\right) > 0.$$

Получив противоречие, мы доказали, что существует такое i, что перенос вершины v_k в сообщество m_i увеличит модулярность, и, как следствие, доказали отсутствие тривиальных сообществ в максимальном по модулярности разбиении простого связного графа.

Из доказанного утверждения получается, что листовая вершина в графе всегда объединяется с сообществом единственного соседа, так как иначе эта вершина лежала бы в тривиальном сообществе. А наличие тривиальных сообществ уменьшает значение модулярности.

Другим важным свойством модулярности является адаптация ее под размер сети. Для начала отметим, что изменение модулярности ΔQ_C при объединении двух сообществ i и j в новое рассчитывается по формуле

$$\Delta Q_C = \frac{k_i^j}{m} - \frac{\Sigma_{\text{tot}}^i \cdot \Sigma_{\text{tot}}^j}{2m^2},$$

где k_i^j – число ребер между сообществами; Σ_{tot}^i – степень i-го сообщества.

Отсюда следует, что при увеличении размеров графа (числа ребер), объединение малых сообществ будет увеличивать модулярность разбиения. Более того, из формулы видно, что при рассмотрении двух связанных сообществ, степень каждого из которых меньше чем $\sqrt{2m}$, их объединение приведет к увеличению модулярности. Теперь можно сформулировать предел разрешения [7] модулярности – одно из ее свойств, которое часто рассматривается как недостаток: в максимальном по модулярности разбиении невзвешенного неориентированного графа не может существовать двух связанных сообществ, степень каждого из которых меньше чем $\sqrt{2m}$. Одним из возможных решений для выделения малых по размеру сообществ является параметризация модулярности:

$$Q(\alpha) = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \alpha\frac{d_i \cdot d_j}{2m}\right) \cdot \delta\left(C_i, C_j\right).$$

Тогда $\Delta Q_C(\alpha)$ будет рассчитываться по формуле

$$\Delta Q_C(\alpha) = \frac{k_i^j}{m} - \alpha\frac{\Sigma_{\text{tot}}^i \cdot \Sigma_{\text{tot}}^j}{2m^2}.$$

Следовательно, минимальный размер двух связанных сообществ для случая с использованием параметризованной модулярности будет равен $\sqrt{\frac{2m}{\alpha}}$.

Отметим, что описанные особенности модулярности находят отражение в алгоритме Блонделя. И хотя разбиение, получаемое алгоритмом Блонделя на заданном графе, зачастую не обладает абсолютно максимальной модулярностью для этого графа, но указанные свойства разбиения при этом выполняются. Это связанно с тем, что отсутствие тривиальных сообществ и существование предела разрешения опираются на величину прироста модулярности, расчет которой является основным шагом работы алгоритма Блонделя.

При рассмотрении графа всей сети объединение вершины с инцидентными ей листьями в одно сообщество оправдано, ведь вершина-сосед является единственным объектом взаимодействия листовой вершины. В силу очень большого размера графа всей сети часто прибегают к анализу некоторых его подграфов. Обычным методом получения подграфа является поиск в ширину от заданной наперед вершины. В то же время при рассмотрении подграфов сетей, в особенности социальных, достаточно часто возникает ситуация, когда некоторая вершина имеет множество соседей-листов. В результате на первом уровне иерархии листовые вершины будут объединены в сообщество с этой вершиной. Отметим, что степень такого

сообщества относительно мала, так как оно содержит в большинстве своем листовые вершины. Из-за этого на следующем уровне иерархии разбиения данное сообщество в силу предела разрешения чаще всего объединяется с другими. При рассмотрении эго-графа пользователя такое сообщество зачастую образуется вокруг его вершины. В то же время наличие таких сообществ не отражает реальной структуры всей сети, большая часть вершин в них будет связана лишь с вершиной пользователя.

Помимо обозначенных выше свойств важным является еще одна особенность, рассмотренная в работе [1]. Заключается она в том, что при выборе, к какому из уже сформированных сообществ добавить вершину, алгоритм неявно основывается на суммарном весе ребер, инцидентных вершинам этих сообществ. Поэтому не редка ситуация, когда какая-то из вершин, имеющая бóльшую степень, относится алгоритмом не к тому сообществу, с которым у нее больше всего общих ребер. Все это вместе приводит к тому, что вершины, инцидентные большому числу листов, которые в силу отсутствия тривиальных сообществ объединяются с ними в одном сообществе, попадают в сообщества с малым суммарным весом. Эти свойства приводят к разбиению, которое можно назвать «сбором мусора». Что в зависимости от исходной задачи анализа сети может быть как удобным, так и неудачным разбиением.

В силу описанных свойств разумным будет рассмотреть некоторые модификации алгоритма Блонделя, которые позволят меньше «собирать мусор».

Тестирование на LFR-моделях и реальных данных

Тестирование на LFR-моделях [8] предполагает создание случайного графа с уже известной структурой сообществ, которая сравнивается со структурой, получаемой тестируемым алгоритмом. Для сравнения разбиений использовалась мера NMI (Normalized Mutual Information, нормализованная взаимная информация), предложенная в качестве меры сходства разбиений графа в 2005 г. [9]. Тестовые LFR-графы были сгенерированы со следующими параметрами:

- число вершин – от 1 000 до 15 000 с шагом 1 000;
- коэффициент смешивания – 0,6;
- максимальная размер сообщества – 50 вершин;
- коэффициент распределения размеров сообществ – 2;
- усреднение на 10 графах.

Как было замечено, алгоритм Блонделя – иерархический, следовательно, в разбиениях больших графов крупные сообщества, расположенные на последнем уровне иерархии, объединяют несколько более малых, полученных на более низких уровнях иерархии. В силу предела разрешения модулярности последний уровень иерархии разбиения алгоритма Блонделя не содержит сообществ меньше определенного размера, даже если последние слабо связаны с остальной сетью. Однако данные сообщества могут быть обнаружены на более низких уровнях иерархии. Так, на рис. 1 видно, что самостоятельное сообщество, отмеченное синей окружностью, на втором уровне иерархии объединяется с центральным сообществом (большое сообщество внизу рисунка).

На LFR-моделях работа с более низкими уровнями иерархии Блонделя показала (рис. 2) наилучшие результаты при малой вариативности весов вершин и сообществ, что, однако, не находит отражения в реальных сетях.

Другой методикой, направленной на уменьшение влияния «сбора мусора» при работе с подграфами, является учет степеней вершин в исходном графе вместо степеней вершин подграфа. Для тестирования данной модификации были получены подграфы сгенерированных LFR-моделей путем обхода в ширину от случайной вершины. Сравнение двух разных подходов – с учетом и без учета реальных степеней вершин, показывает весомое преимущество рассматриваемой модификации (рис. 3).

На практике же при учете реальных степеней требуется использовать масштабируемую модификацию модулярности. Так, при работе с эго-графом (графом друзей определенного пользователя, включая его самого) сети Instagram без масштабирующего коэффициента было получено разбиение, представленное на рис. 4.

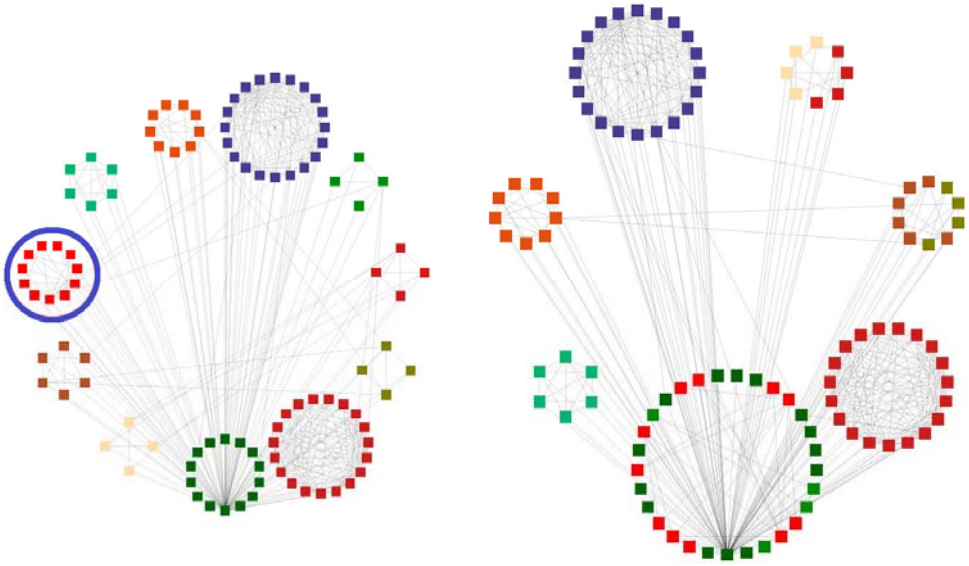

Рис. 1. Первый (слева) и второй (справа) уровни иерархии Блонделя

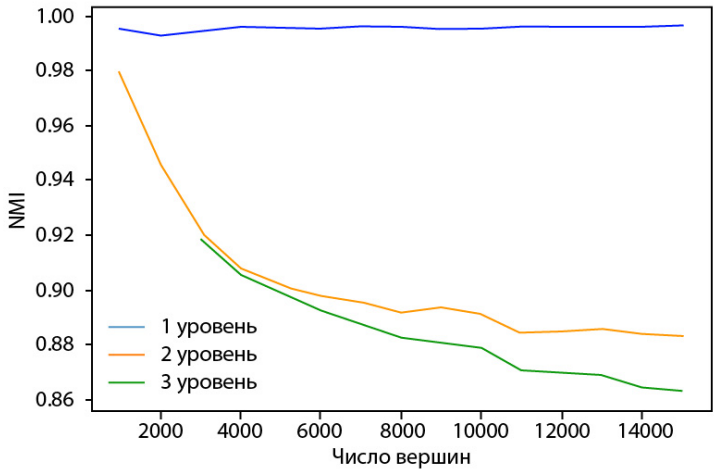

Рис. 2. NMI на различных уровнях иерархии Блонделя

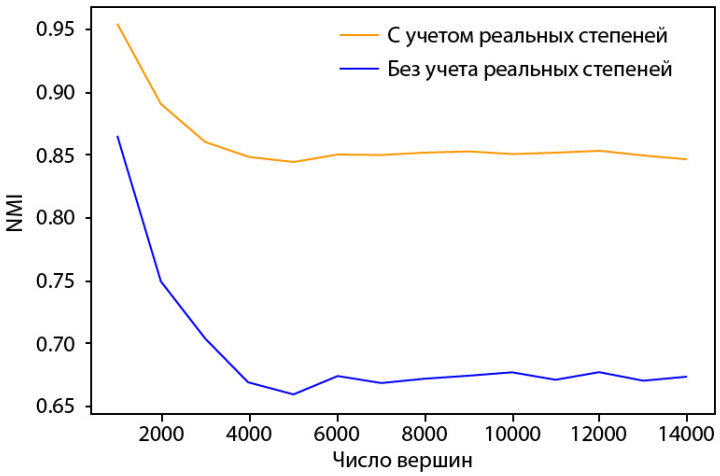

Рис. 3. Сравнение модификации, учитывающей реальный вес вершины, с оригинальным алгоритмом

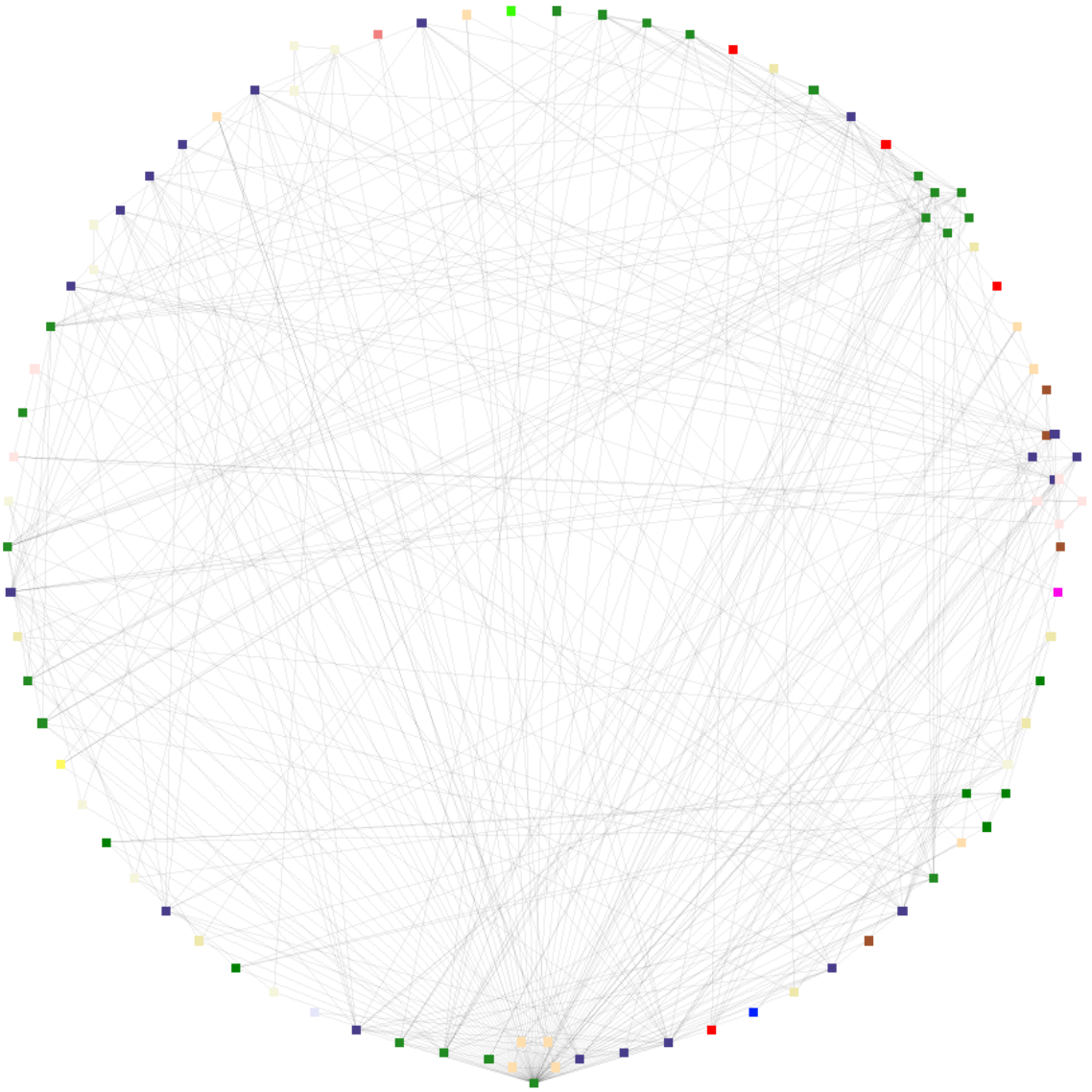

Рис. 4. Результат работы алгоритма Блонделя с учетом реальных весов на реальной сети

Как видно из рисунка, бóльшая часть вершин лежит в тривиальных сообществах. Это связано с большой степенью вершин при малых размерах графа. В данном случае для получения более содержательных разбиений можно искусственно ослаблять критерий прироста модулярности. Для этого воспользуемся введенной нами ранее параметризованной модулярностью. Уменьшая параметр α, получим различные по своей структуре разбиения. Так, при $\alpha = 0,05$ получаем разбиение, представленное на рис. 5.

Следует отметить, что благодаря учету реальных степеней удается избежать «сбора мусора», и вокруг вершины пользователя не образуется «мусорное сообщество», а, наоборот, она добавляется в сообщество, с которым имеет максимальное взаимодействие.

Заключение

В данной статье авторы рассмотрели один из популярных иерархических агломеративных алгоритмов выделения сообществ – алгоритм Блонделя. Подробно описаны и доказаны

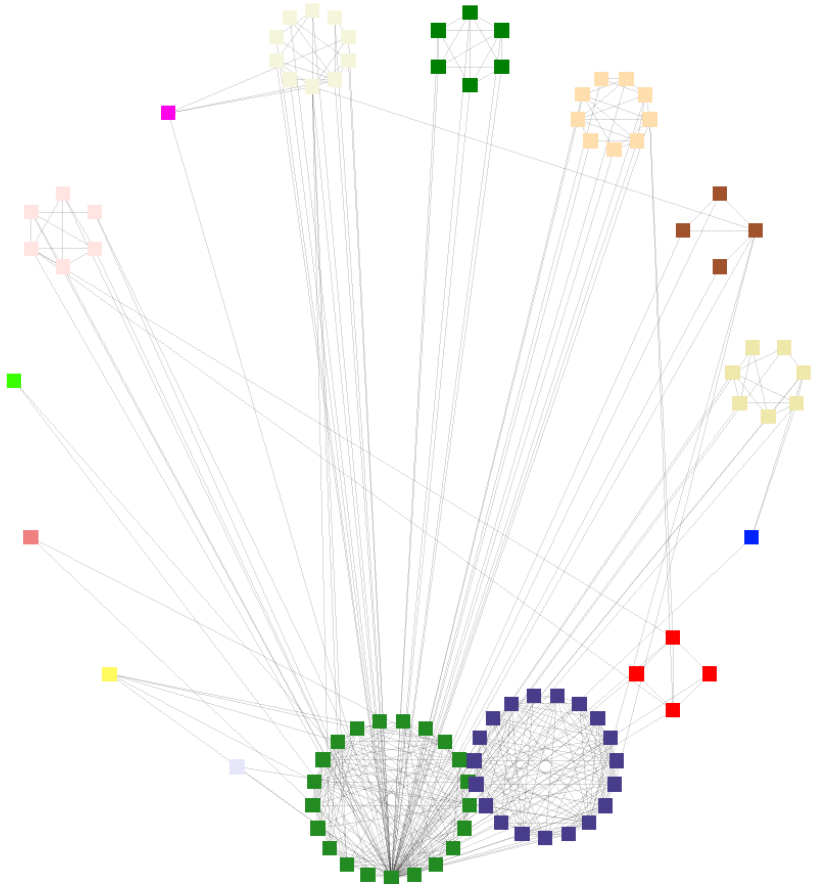

Рис. 5. Результат работы алгоритма Блонделя
с учетом реальных весов при α = 0,05

некоторые свойства этого алгоритма и получаемого после его применения разбиения графа на сообщества, в том числе указаны качественные недостатки, которые могут быть критичными в зависимости от исходной задачи анализа сети. Сделаны предложения о возможных модификациях, которые помогают получить иные, репрезентативные результаты. В частности, рассмотрен вариант учета реальных степеней вершин в подграфе и параметризованной модулярности. Данные вариации протестированы не только на сгенерированных графах LFR-модели, но и на реальных данных, полученных из социальных сетей. Считаем целесообразным использовать предложенные подходы при анализе сетей взаимодействующих объектов наряду с классическими алгоритмами.

Список литературы

1. *Girvan M.*, *Newman M. E. J.* Community structure in social and biological networks // Proc. Natl. Acad. Sci. USA. 2002. Vol. 99, № 12. P. 7821–7826.

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

3. *Newman M. E. J.* Modularity and community structure in networks // Proc. Natl. Acad. Sci. USA. 2006. Vol. 103, № 23. P. 8577–8582.

4. *Brandes U.*, *Delling D.*, *Gaertler M.*, *Goerke R.*, *Hoefer M.*, *Nikoloski Z.*, *Wagner D.* Maximizing Modularity is hard // arXiv:physics/0608255. 2006.

5. *Blondel V.*, *Guillaume J.*, *Lambiotte R.*, *Lefebvre E.* Fast unfolding of communities in large networks // Journal of Statistical Mechanics: Theory and Experiment. 2008. Vol. 10. P. 10008.

6. *Dugue N.*, *Perez A.* Directed Louvain: maximizing modularity in directed networks. Research Report. Uni. d`Orleans Press, 2015. <hal-01231784>

7. *Fortunato S.*, *Barthélemy M.* Resolution limit in community detection // Proc. Natl. Acad. Sci. USA. 2007. Vol. 104. № 36.

8. *Lancichinetti A.*, *Fortunato S.* Benchmark graphs for testing community detection algorithms // Physical Review. E 78. 2008. P. 046110.

9. *Danon L.*, *Duch J.*, *Diaz-Guilera A.*, *Arenas A*. Comparing community structure identification // arXiv:cond-mat/0505245. 2005.

Материал поступил в редколлегию 24.07.2017

A. O. Orlov, **A. A. Chepovskiy**

National Research University – Higher School of Economics
20 Myasnitskaya St., *Moscow*, *101000*, *Russian Federation*

aachepovsky@hse.ru

ABOUT MODULARITY PROPERTIES AND ACTUAL ADJUSTMENTS OF THE BLONDEL ALGORITHM

One of the tasks related to the study of the of complex networks is the task of revealing communities structure – splitting all vertices into groups (communities), so that the vertices of each group are more closely related to each other than to the rest of the graph. A popular algorithm for detecting communities is the Blondel, based on the maximization of Newman-Girvan modularity, a common criterion for assessing the quality of community divisions. This article is devoted to the analysis of its features and work results, as well as possible modifications. The test results are analyzed both on the generated graphs and on real data.

Keywords: graph structure, social network analysis, community detection, big data.

For citation:

Orlov A. O., Chepovskiy A. A. About Modularity Properties and Actual Adjustments of the Blondel Algorithm. *Vestnik NSU. Series*: *Information Technologies*, 2017, vol. 15, no. 3, p. 64–73. (In Russ.)