Наука и Образование
МГТУ им. Н.Э. Баумана
Сетевое научное издание
ISSN 1994-0408

Ссылка на статью:
// Наука и Образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2017. №4. С. 137–152.
DOI: 10.7463/0417.0001133
Представлена в редакцию: 19.03.2017

УДК 519.178+51-77

Современные методы выделения сообществ в социальных сетях

Чесноков В. О.[1,*], **Ключарёв П. Г.**[1]

[*]v.o.chesnokov@yandex.ru

[1]МГТУ им. Н.Э. Баумана, Москва, Россия

В статье приведен обзор основных подходов к выделению сообществ в социальных сетях. Распространенный способ решения этой задачи заключается в применении одного из алгоритмов кластеризации графов к социальной сети. Однако социальные сети имеют ряд особенностей: сообщества могут пересекаться и иметь иерархическую структуру, а с вершинами обычно связано множество атрибутов. При этом информация об атрибутах вершин может частично отсутствовать. Поэтому стоит использовать, алгоритмы разработанные с учетом этих особенностей. Однако даже самые продвинутые методы не могут дать результаты, близкие к эталонным покрытиям, созданными человеком-экспертом. В данной области стоит ожидать совершенствования существующих алгоритмов и появления новых.

Ключевые слова: выделение сообществ; социальный граф; социальная сеть

Введение

Хорошо известно, что многие естественные сети имеют неоднородную структуру. Между некоторыми группами вершин больше связей, чем в среднем в графе — такие «уплотнения» называют кластерами или сообществами. Социальные сети, т.е. графы в которых вершины соответствуют людям, а ребра — некоторым связям между ними, не являются исключением.

К сожалению, в настоящий момент не существует общепринятого определения сообщества [1, 2, 3]. Более того, определение сообщества может варьироваться в зависимости от предметной области и даже решаемой задачи. Поэтому будем называть сообществом $C_i(V_i, E_i)$ любой вершинно-порожденный подграф социального графа $G(V, E)$:

$$V_i \subseteq V, \qquad E_i \subseteq E : \forall \{v, w\} \in E_i \Longrightarrow v \in V_i, w \in V_i.$$

Множество сообществ графа $\mathcal{C} = \{C_i\}$ будем назвать покрытием. В некоторых графах сообщества могут быть неоднородными [2, 4]: более плотную, центральную часть называют ядром, а прочие вершины составляют периферию.

Отличительная особенность сообществ в социальных сетях заключается в том, что некоторые вершины могут принадлежать нескольким сообществам, а некоторые — не состоять

ни в одном. Кроме того, зачастую с вершинами связано множество атрибутов, таких как демографические характеристики, социальный статус и т.п.

На данный момент было предложено очень много алгоритмов выделения сообществ, и полный их обзор будет весьма трудоемким и нецелесообразным. Рассмотрим наиболее значимые из них.

1. Кластеризация графов

Довольно часто исследователи для в качестве метода выделения сообществ используют один из алгоритмов кластеризации графа [5, 6, 7]. Один из самых известных и часто используемых методов — максимизация модулярности. Модулярность графа — мера, показывающая насколько ребер внутри групп вершин больше, чем снаружи [8]:

$$Q = \sum_i (e_{ii} - a_i^2),$$

где e_{ij} — доля ребер, соединяющих сообщества i и j, а $a_i = \sum_j e_{ij}$. Большинство вариаций алгоритма реализованы с использованием жадного подхода. Существуют алгоритмы, основанные и на оптимизации других внутренних мер графа [2]. Примерами таких мер могут служить плотность, которая может быть вычислена как

$$D = \frac{2|E|}{|V|(|V|-1)},$$

и проводимость [9], которая также называется коэффициентом реберного расширения или константой Чигера:

$$h(G) = \min_{\{S \subset V | 0 < |S| \leq \frac{|V|}{2}\}} \frac{|\partial S|}{|S|},$$

где ∂S — множество ребер, каждое из которых инцидентно ровно одной вершине из некоторого подмножества $S \subset V$ [10].

Отличительной особенностью методов, основанных на случайных блужданиях, является тот факт, что им не требуется единовременное знание о всей структуре графа, а только о некоторой его части. Примером такого алгоритма может служить Infomap [11]. В его основе лежит оптимизация способа кодирования узлов графа на пути случайного блуждания с помощью кода Хаффмана. Чтобы код всего пути был минимален, кластеры кодируются отдельно и для каждого из них вводят специальный код «выхода из кластера».

В то время как большинство алгоритмов выделения сообществ имеют полиномиальную сложность относительно количества вершин [2], методы, основанные на переносе меток, являются одними из немногих, имеющих квазилинейную сложность. Первый алгоритм из этого семейства был предложен Рагхаваном, Альберт и Кумарой в работе [12]. В нем каждой вершине x присваивают случайную метку $c(x, 0)$. Затем начинается итеративный процесс, в котором для каждой вершины x метку заменяют на ту, которую имеют большинство соседей:

$$c(x, t+1) = \text{argmax}\, |V_x(c)|,$$

где $V_x(c) = \{v \in N(x) | c(v) = c\}$, а $N(x)$ — множество соседей вершины x. После завершения этого процесса группы вершин с одинаковыми метками составляют кластеры:

$$C_c = \{v \in V | c(v) = c\}.$$

В целом в большом множестве алгоритмов выделения сообществ можно выделить пять основных классов методов [2]:

•

– основанные на оптимизации некоторой меры — например, максимизация модулярности;

– объединяющие вершины в кластеры по некоторой мере схожести — например, k-средних или основанные на анализе спектра графа;

– основанные на обнаружении подграфов с заданными свойствам — например, Clique Percolation [13], который основан на поиске пересекающихся клик, и SCAN [14], в котором идет обнаружение вершин, принадлежащих ядрам сообществ;

– разбивающие граф путем удаления ребер — например, алгоритм последовательного удаления ребер с большой промежуточностью [5], которая определяется частотой вхождения ребра в кратчайшие пути между парами вершин;

– основанные на некоторой вероятностной модели или модели динамического процесса — например, Infomap или перенос меток.

2. Выделение пересекающихся сообществ

В последние годы исследователи пришли к выводу, что классические алгоритмы кластеризации неприменимы к социальным сетям [2, 3, 15, 16]. Каждый человек естественным образом состоит в нескольких сообществах, соответствующих его сферам деятельности, поэтому при исследовании социальных графов стоит использовать алгоритмы, обнаруживающие пересекающиеся и иерархические сообщества. Лишь некоторые алгоритмы выделяют сообщества с учетом этого обстоятельства [16]. Достаточно большое количество таких методов являются доработанными методами кластеризации графов [17, 18, 19].

Некоторые методы основаны на кластеризации ребер, а не вершин. Примером такого алгоритма может служить метод, предложенный в работе [20]. В начале каждое ребро состоит в своем кластере. Затем на каждой итерации определяют два наиболее схожих ребра, и соответствующие им кластеры объединяют. В качестве меры схожести ребер $\{i, k\}$ и $\{j, k\}$ используют индекс Жаккара:

$$J = \frac{|N_+(i) \cap N_+(j)|}{|N_+(i) \cup N_+(j)|},$$

где $N_+(x) = N(x) \cup \{x\}$. Итерации завершаются, когда остался ровно один кластер, содержащий все ребра. Из всех возможных разбиений на промежуточных этапах, выбирают то, в

котором достигнуто наибольшее значение меры плотности разбиения:

$$D_{\mathcal{P}} = \frac{2}{|E|} \sum_{P \in \mathcal{P}} |P| \frac{|P| - (n_P - 1)}{(n_P - 2)(n_P - 1)},$$

где $\mathcal{P} = \bigcup P_i$ — разбиение множества ребер E, а $n_P = |\bigcup_{\{i,j\} \in P} \{i, j\}|$. Вершина лежит в нескольких сообществах, если ребра, инцидентные ей, попали в разные кластеры.

Часть алгоритмов основана на подходе локального расширения сообщества из одной вершины [16]. Например, в алгоритме LFM [21] сообщество «растет» из случайно выбранной вершины. Производится обход всех соседей вершины, и в сообщество добавляют ту, которая имеет наибольшее значение фитнес-функции $f_{V_i}(x) = f_{V_i \cup \{x\}} - f_{V_i \setminus \{x\}}$, где

$$f_A = \frac{k_{\text{in}}^A}{(k_{\text{in}}^A + k_{\text{out}}^A)^\alpha}$$

— фитнес-функция множества $A \in V$; k_{in}^A и k_{out}^A — суммарные внутренние и внешние степени вершин; α — параметр алгоритма, действительное число от 0 до 1. Затем из сообщества удаляют все вершины с отрицательным значением фитнес-функции, и процедура повторяется. Формирование сообщества завершается, когда все соседние вершины имеют отрицательное значение фитнес-функции. После получения сообщества выбирается вершина графа, не состоящая в сообществах, и процедура повторяется.

В некоторых подходах используются вектора принадлежности. Каждой вершине соответствует один вектор, в котором хранятся степени принадлежности каждому из сообществ. Например, алгоритмы в работах [22] и [23] основаны на неотрицательном матричном разложении (NMF) — технике, заимствованной из машинного обучения. Однако такие методы обычно требуют знания количества сообществ и имеют высокую вычислительную сложность.

Тем не менее, алгоритмы, основанные на переносе меток, избавлены от этих недостатков. Алгоритм SPLA [24], который является доработанным алгоритмом переноса меток из работы [12], построен по модели «слушающих» и «говорящих» узлов, при каждая вершина «помнит», какие метки у нее были с какой частотой — эта информация и будет вектором принадлежности. Заслуживает упоминания и доработанный вариант этого алгоритма — EgoLP [25], в котором сообщества графа определяются на основе покрытий $\mathcal{C}_x$, $x \in V$, вершинно-порожденных подграфов $G_x(N(x), N_E(x))$, где $N_E(x)$ — множество ребер, соединяющих вершины из $N(x)$. Покрытия $\mathcal{C}_x$ получают с помощью алгоритма SPLA.

Среди методов, выделяющих пересекающиеся сообщества, примечательны алгоритмы, разработанные группой Лесковица. Они предложили модель присоединения, в которой считается, что формирование графа происходит под влиянием внешних факторов, т.е. сообщества образуются из-за наличия общих атрибутов у вершин. Пусть имеется неотрицательная матрица $F = \{f_{u,C}\}$, в которой $f_{u,C}$ — вес между вершины $u \in V$ в сообществе C. Алгоритм BigCLAM [26] строит граф $G(V, E)$, создавая ребра по следующему правилу:

$$p(u, v) = 1 - \exp(-F_u \cdot F_v^T),$$

где $p(u,v)$ — вероятность создания ребра между вершинами u и v, а F_u — вектор весов для вершины u. Структура сообществ графа определяется путем нахождения такой матрицы F, при которой достигается максимум правдоподобия $l(F) = \log P(G|F)$:

$$\widehat{F} = \text{argmax}\, l(F),$$

где функция правдоподобия вычисляется как

$$l(F) = \mathcal{L}_G = \sum_{(u,v)\in E} \log p(u,v) - \sum_{(u,v)\notin E} F_u F_v^T.$$

В работах [26, 27] показано превосходство алгоритмов AGM-fit и его варианта для больших сетей BigCLAM, над такими алгоритмами как Clique Percolation [13], Mixed-Membership Stochastic Block Models [28] и основанных на кластеризации ребер [20].

3. Выделение сообществ с использованием атрибутов вершин

В алгоритмах, описанных ранее, информация об атрибутах вершин не используется для получения сообществ. Иногда атрибуты могут использоваться в дальнейшем для именования сообществ, однако часто эта операция осуществляется человеком вручную.

В то же время социальные графы характеризуются тем, что о пользователях известно достаточно много данных — сведения об образовании, местах работы, прохождения воинской службы, членстве в виртуальных сообществах по интересам (которые часто являются отражением реальных) и т.п. Сама социальная сеть побуждает пользователя указать как можно данных о себе. Существуют алгоритмы выделения сообществ, которые концентрируются на таких данных.

Примером могут служить алгоритмы, основанные на латентном размещении Дирихле [29], которое изначально предназначалось для кластеризации текстов. В этой модели документ представляет собой смесь из неявных тем, которые в свою очередь определены некоторым распределением вероятности появления слов. Путем порождающего процесса для этой модели определяются наиболее правдоподобные соответствия тем и документов. С точки зрения теории графов документы можно рассматривать как вершины, темы — как сообщества, а слова соответствуют атрибутам вершин.

Другим примером может служить алгоритм CODICIL [30]. На первом этапе этого алгоритма формируют множество ребер схожести вершин E_t по атрибутам из множества всех атрибутов T. Для каждой вершины u выбирают k ее соседей, с которыми функция схожести Sim множеств атрибутов имеет наибольшее значение. В CODICIL она вычислялась как косинус TF-IDF векторов атрибутов двух вершин:

$$\text{Sim} = \frac{tf\text{-}idf(u) \cdot tf\text{-}idf(v)}{||tf\text{-}idf(u)||_2 \cdot ||tf\text{-}idf(v)||_2},$$

где i-я координата вектора $tf\text{-}idf(u)$ вычисляется частоте вхождения i-го атрибута $tf_i(u)$ как

$$tf\text{-}idf(u,i) = \sqrt{tf_i(u)} \log\left(1 + \frac{|T|}{\sum_{j=1}^{|T|} tf_j(u)}\right).$$

Затем множества E_t и E объединяют в множество E^* и выбирают из него множество $E^*_{\text{sample}} \subset E^*$ такое, что $|E^*_{\text{sample}}| << |E|$. Для этого для каждой вершины i из E^* берут $\sqrt{|N(i)|}$ ребер, которые соединяют ее с наиболее похожими соседями, т.е. у которых наибольшее значение индекса Жаккара для множеств $N(i)$ и $N(j)$. Наконец, к графу $G(V, E^*_{\text{sample}})$ применяют один из алгоритмов кластеризации графов, результат которого будет итогом основного алгоритма.

В работе [31] предложен метод, основанный на переносе меток. В качестве меток используют не случайные значения, а имеющиеся атрибуты вершин, и обновляют метки по мажоритарному правилу. За счет того, что многие вершины имеют одинаковые метки при инициализации алгоритма, он имеет низкую вычислительную сложность.

Алгоритмы, предназначенные для выделения сообществ в графах с атрибутами вершин можно разделить на следующие категории [32]:

•

– методы, переводящие граф с атрибутами вершин во взвешенный граф, в котором вес ребра вычисляется в зависимости от схожести инцидентных ему вершин;

– методы, строящие функцию расстояния между вершинами на основе структуры графа и схожести вершин с последующим применением алгоритма кластеризации;

– методы, основанные на случайных блужданиях по смешанному графу (атрибуты становятся вершинами);

– методы, основанные на вероятностных моделях;

– методы, основанные на разбиении пространства атрибутов;

– прочие методы.

Стоит отметить, что многие алгоритмы основаны на предположении, что все атрибуты вершины в сообществе схожи [32, 33, 34]. Как было показано в [35], это предположение неверно для социальных сетей. Например, одноклассники могут поступить в разные вузы и иметь разное место работы.

4. Выделение пересекающихся сообществ по атрибутам вершин

Лишь недавно начали появляться алгоритмы, в которых используется информация и о ребрах графа, и об атрибутах вершин. При этом только немногие из них позволяют выделять пересекающиеся сообщества.

В работе [36] предложен алгоритм, в котором сообщества выделяются с помощью разбиения пространства атрибутов $\mathcal{A} = \{A_1,\ A_2,\ \ldots, A_m\}$, $A_i \subseteq \mathbb{R}$, на ячейки. В модели данного алгоритма i-й атрибут принимает одно из значений из множества A_i. Ячейка пространства представляет собой некоторый полуинтервал одного из множеств A_i. Вместо того, чтобы

рассматривать все подпространства, в алгоритме рассматриваются только «интересные», обладающие низкой энтропией атрибутов. Сообщества получают путем объединения соседних ячеек с высокой плотностью и связностью, которая вычисляется для ячейки c как сумма весов случайных блужданий по всем парам вершин:

$$D(c) = \frac{\sum_{v_i,v_j \in c} \widetilde{q}_{v_i,v_j}}{|c|},$$

где $\widetilde{Q} = \{\widetilde{q}_{v_i,v_j}\}$ — нормализованная матрица весов случайных блужданий между вершинами v_i и v_j. Стоит отметить, что данный алгоритм требует знания о природе атрибутов и их фильтрацию, поскольку использует многомерные пространства атрибутов. Кроме того, в худшем случае он имеет достаточно высокую вычислительную сложность — $O(|V|^2 \cdot 2^{|\mathcal{A}|})$.

Алгоритм EDCAR [37] тоже использует разбиение множества атрибутов на подпространства и основан на эвристике поиска клик. Тем не менее он также требует большого количество вычислений.

Несколько алгоритмов основаны на тематическом моделировании, например, Block-LDA [38], в котором вероятностная модель LDA [29] дополнена моделью связей между документами (вершинами). Однако большинство подобных алгоритмов работают весьма медленно и плохо масштабируются.

Алгоритм CESNA, предложенный группой Лесковица в [15], основан на вероятностной генеративной модели. В нем была улучшена модель алгоритма BigCLAM, чтобы учитывать и атрибуты вершин. Для этого вводится матрица $W = \{w_{t,C}\}$, в которой $w_{t,C}$ показывает релевантность атрибута t сообществу C. Вместо функции $l(F)$ теперь максимизируется $l(F,W) = \mathcal{L}_G + \mathcal{L}_X$, где

$$\mathcal{L}_{\mathcal{X}} = \sum_{u,t} \left(x_{u,t} \log q_{u,t} + (1 - x_{u,t}) \log(1 - q_{u,t})\right),$$

$$q_{u,t} = \frac{1}{1 + \exp(-\sum_C w_{t,C} f_{u,C})},$$

а $x_{u,t}$ — характеристическая функция наличия у вершины u атрибута t. Авторы показали превосходство этого метода над такими алгоритмами, как Block-LDA, CODICIL и EDCAR на выборках данных из онлайновых социальных сетей. Более того, разработанный алгоритм хорошо масштабируется и имеет почти линейную сложность.

В недавней работе [39] был предложен алгоритм, в котором задача выделения сообществ формулируется как задача об неотрицательном матричном разложении:

$$\min_{Q,S,H \geq 0} d(A||QSQ^T) + d(X||QH),$$

где $d(A||B)$ — мера схожести матриц A и B, которая вычисляется как

$$d(A||B) = \sum_{i \neq j} \left(a_{i,j} \log \frac{a_{i,j}}{b_{i,j}} + a_{i,j} - b_{i,j}\right),$$

где $q_{C,u}$ — вероятность того, что ребро в сообществе C инцидентно вершине u; $s_{C,C}$ — вероятность того, что ребро лежит в сообществе C; $h_{C,t}$ — вероятность наличия атрибута t у вершины из сообщества C. Алгоритм дает сопоставимые с CESNA результаты по качеству выделения сообществ, но имеет при этом более высокую вычислительную сложность.

Алгоритм FCAN [40] также использует неотрицательное матричное разложение и основан на максимизации меры r релевантности содержимого вершин в кластере, которая для вершин v_i и v_j имеет значение

$$r(v_i, v_j) = \frac{\sum_{p=1}^{|T|} \sum_{q=1}^{|T|} \text{strength}(val_{i,p}, val_{j,q})}{|T|^2},$$

где strength — некоторая нормализованная функция схожести; $val_{i,p}$ — значение p-го атрибута для вершины v_i; T — множество всех атрибутов. Он превосходит CESNA на синтетических наборах данных, однако на выборах данных из социальных сетей Facebook и Twitter алгоритм CESNA дает более высокие значения меры взаимной информации (NMI) и доли правильных ответов.

Пока не известны алгоритмы, превосходящие результаты алгоритма CESNA на реальных данных. Однако получение покрытия, совпадающего с размеченным человеком вручную, все еще является трудной задачей. В связи с этим требуется совершенствование существующих алгоритмов.

5. Устойчивость к отсутствию атрибутов

Почти все алгоритмы выделения сообществ подразумевают, что предоставлена полная информация о графе и атрибутах вершин [31]. Исключением является алгоритм CESNA, создатели которого исследовали его устойчивость к частичному отсутствию ребер. Как было показано ранее, не все пользователи социальной сети выставляют информацию о себе на публичное обозрение, и она может быть скрыта настройками приватности, что затрудняет ее анализ.

В связи этим возникает необходимость разработки такого алгоритма выделения сообществ, который будет устойчив к частичному отсутствию атрибутов вершин. Данная проблема тесно связана с проблемой определения неуказанной или отсутствующей информации.

Заключение

Классический подход к выделению сообществ заключается в применении к социальному графу одного из алгоритмов кластеризации. Наиболее популярны методы, основанные на максимизации модулярности.

Однако такой метод не учитывает две важные особенности социального графа: наличие атрибутов вершин и возможность пересечения сообществ. Лишь недавно стали появляться методы, которые используют как информацию о структуре графа, так и сведения об атрибутах

вершин и позволяют выделять пересекающиеся сообщества. На данный момент наиболее точные результаты дает алгоритм CESNA. Хотя он может и проигрывать по некоторым показателям другим алгоритмам на синтетических графах, он превосходит аналоги на реальных выборках данных из социальных сетей.

Стоит отметить, что в исследованиях практически не уделяется внимание проблеме отсутствующих атрибутов вершин. В социальных сетях не всегда доступна полная информация о профиле пользователя: они могут быть скрыты настройками приватности или просто не указаны.

К сожалению, на данный момент самые лучшие алгоритмы далеки от совершенства. Если оценивать качество их работы на графах с известными покрытиями, созданными человеком вручную, то они показывают результаты, далекие от точного совпадения. Поэтому стоит ожидать дальнейшего развития этой области и появления новых алгоритмов.

Данная работа проведена при поддержке гранта РФФИ No 16-29-09517 офи_м «Методы и алгоритмы выявления сообществ и организации информационного противоборства в социальных сетях на основе байесовских и теоретико-игровых подходов с использованием графовых и фрактальных моделей».

Список литературы

1. Fortunato S.. Community detection in graphs // Physics Reports. 2010. Vol. 486, no. 3-5. P. 75–174.
2. Papadopoulos S., Kompatsiaris Y., Vakali A., Spyridonos P. Community detection in Social Media // Data Mining and Knowledge Discovery. 2012. Vol. 24, no. 3. P. 515–554.
3. Коршунов А.В. Задачи и методы определения атрибутов пользователей социальных сетей // Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», RCDL'2013. 2013.
4. 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 // CoRR. 2008. Vol. abs/0810.1355.
5. Newman M.E.J., Girvan M. Finding and evaluating community structure in networks // Physical Review E. 2004. Vol. 69, iss. 2. Art. no. 026113 (15 pages). DOI: 10.1103/PhysRevE.69.026113
6. Schaeffer S.E.. Graph clustering // Computer Science Review. 2007. Vol. 1, no. 1. P. 27–64.
7. Scott J.. Social network analysis: A handbook. 2nd ed. London: SAGE, 2000.
8. Newman M.E.J. Fast algorithm for detecting community structure in networks // Physical Review E. 2004. Vol. 69, iss. 6. Art.no. 066133 (5 pages). DOI: 10.1103/PhysRevE.69.066133
9. Kannan R., Vempala S., Vetta. A. On clusterings: Good, bad and spectral // Journal of the ACM. 2004. Vol. 51, no. 3. P. 497-515.

10. Hoory S., Linial N., Wigderson A. Expander graphs and their applications // Bulletin of American Mathematical Society. 2006. Vol. 43, no. 4. P. 439-561. DOI: 10.1090/S0273-0979-06-01126-8
11. Rosvall M., Bergstrom C.T. Maps of random walks on complex networks reveal community structure // Proceedings of the National Academy of Sciences. 2008. Vol. 105, no. 4. P. 1118–1123.
12. 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. Art. no. 036106 (11 pages).
13. Palla G., Derenyi I., Farkas I., Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society // Nature. 2005. Vol. 435, no. 7043. P. 814–818.
14. Xu X., Yuruk N., Feng Z., Schweiger T.A.J. SCAN: A Structural Clustering Algorithm for Networks // Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'07. New York, NY, USA: ACM, 2007. P. 824–833.
15. Yang J., McAuley J.J., Leskovec J. Community Detection in Networks with Node Attributes // CoRR. 2014. Vol. abs/1401.7267.
16. Xie J., Kelley S., Szymanski B.K. Overlapping Community Detection in Networks: The State-of-the-art and Comparative Study // ACM Comput. Surv. 2013. Vol. 45, no. 4. P. 43:1–43:35.
17. Nicosia V., Mangioni G., Carchiolo V., Malgeri M. Extending the definition of modularity to directed graphs with overlapping communities // Journal of Statistical Mechanics: Theory and Experiment. 2009. Vol. 2009, no. 3. Art.ṅo. P03024.
18. Lazar A., Abel D., Vicsek T. Modularity measure of networks with overlapping communities // EPL (Europhysics Letters). 2010. Vol. 90, no. 1. Art.ṅo. 18001. 10.1209/0295-5075/90/18001
19. Бузун Н.О., Коршунов А.В. Выявление пересекающихся сообществ в социальных сетях // Доклады Всероссийской научной конференции «Анализ изображений, сетей и текстов», АИСТ'2012. 2012.
20. Ahn Y.-Y.Y., Bagrow J.P., Lehmann S. Link communities reveal multiscale complexity in networks // Nature. 2010. Vol. 466, no. 7307. P. 761–764.
21. Lancichinetti A., Fortunato S., Kertesz J. Detecting the overlapping and hierarchical community structure in complex networks // New Journal of Physics. 2009. Vol. 11, no. 3. Art.ṅo. 033015.
22. Zhao K., Zhang S., Pan Q. Fuzzy analysis for overlapping community structure of complex network // Control and Decision Conference, CCDC. IEEE Computer Society, 2010. P. 3976–3981.
23. Psorakis I., Roberts S., Ebden M., Sheldon B. Overlapping community detection using Bayesian non-negative matrix factorization // Physical Review E. 2011. Vol. 83, iss. 6. Art. no. 066114 (9 pages). DOI: 10.1103/PhysRevE.83.066114

24. Xie J., Szymanski B.K., Liu X. SLPA: Uncovering Overlapping Communities in Social Networks via a Speaker-Listener Interaction Dynamic Process // Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops, ICDMW '11. Washington, DC, USA: IEEE Computer Society, 2011. P. 344–349.
25. Buzun N, Korshunov A., Avanesov V., Filonenko I., Kozlov I., Turdakov D., Kim H. EgoLP: Fast and Distributed Community Detection in Billion-Node Social Networks // 2014 IEEE International Conference on Data Mining Workshop. 2014. P. 533–540.
26. Yang J., Leskovec J. Overlapping Community Detection at Scale: A Nonnegative Matrix Factorization Approach // Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13. New York, NY, USA: ACM, 2013. P. 587–596.
27. Yang J., Leskovec J. Community-Affiliation Graph Model for Overlapping Network Community Detection // 12th IEEE International Conference on Data Mining, ICDM-2012. Brussels, Belgium, December 10-13, 2012. P. 1170–1175.
28. Airoldi E.M., Blei D.M., Fienberg S.E., Xing E.P. Mixed Membership Stochastic Blockmodels // Journal of Machine Learning Research. 2008. Vol. 9. P. 1981–2014.
29. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. Vol. 3. P. 993–1022.
30. Ruan Y., Fuhry D., Parthasarathy S. Efficient Community Detection in Large Networks Using Content and Links // Proceedings of the 22nd International Conference on World Wide Web, WWW '13. New York, NY, USA: ACM, 2013. P. 1089–1098.
31. Чесноков В.О., Ключарёв П.Г. Выделение сообществ в социальных графах по множеству признаков с частичной информацией // Наука и Образование. МГТУ им. Н.Э.Баумана. Электрон. журн. 2015. № 9. С. 188-199. DOI: 10.7463/0915.0811704
32. Bothorel C., Cruz J.D,Magnani M., Micenkova B. Clustering attributed graphs: Models, measures and methods // Network Science. 2015. Vol. 3, no. 3. P. 408–444.
33. Neville J., Adler M., Jensen D. Clustering relational data using attribute and link information // In Proceedings of the Text Mining and Link Analysis Workshop, 18th International Joint Conference on Artificial Intelligence. 2003. P. 9–15.
34. Zhou Y., Cheng H., Yu J.X. Graph Clustering Based on Structural/Attribute Similarities // Proc. VLDB Endow. 2009. Vol. 2, no. 1. P. 718–729.
35. Li R., Wang C., Chang K.C.-C. User Profiling in an Ego Network: Co-profiling Attributes and Relationships // Proceedings of the 23rd International Conference on World Wide Web. Seoul, Korea: ACM, 2014. P. 819–830.
36. Huang X., Cheng H., Yu J.X. Dense Community Detection in Multi-valued Attributed Networks // Inf. Sci. 2015. Vol. 314, no. C. P. 77–99.
37. Gunnemann S., Boden B., Farber I., Seidl T. Efficient Mining of Combined Subspace and Subgraph Clusters in Graphs with Feature Vectors // Advances in Knowledge Discovery and

Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part I. Springer Berlin Heidelberg, 2013. P. 261–275.

38. Balasubramanyan R., Cohen W.W. Block-LDA: Jointly Modeling Entity-Annotated Text and Entity-Entity Links // Handbook of Mixed Membership Models and Their Applications. 2014. P. 255–273.
39. Nguyen H.T., Dinh T.N. Unveiling the structure of multi-attributed networks via joint non-negative matrix factorization // Military Communications Conference, MILCOM 2015. IEEE. 2015. P. 1379–1384.
40. Hu L., Chan K. C.C. Fuzzy Clustering in a Complex Network Based on Content Relevance and Link Structures // IEEE Transactions on Fuzzy Systems. 2016. Vol. 24, no. 2. P. 456–470.

Science & Education of the Bauman MSTU
Electronic journal
ISSN 1994-0408

Science and Education of the Bauman MSTU, 2017, no. 4, pp. 137–152.
DOI: **10.7463/0417.0001133**
Received: 19.03.2017

Modern community detection methods in social networks

Chesnokov V. O.[1,*], **Klucharev P. G.**[1]

[*]v.o.chesnokov@yandex.ru

[1]Bauman Moscow State Technical University, Russia

Keywords: community detection, social graph, social network

Social network structure is not homogeneous. Groups of vertices which have a lot of links between them are called communities. A survey of algorithms discovering such groups is presented in the article.

A popular approach to community detection is to use an graph clustering algorithm. Methods based on inner metric optimization are common. 5 groups of algorithms are listed: based on optimization, joining vertices into clusters by some closeness measure, special subgraphs discovery, partitioning graph by deleting edges, and based on a dynamic process or generative model.

Overlapping community detection algorithms are usually just modified graph clustering algorithms. Other approaches do exist, e.g. ones based on edges clustering or constructing communities around randomly chosen vertices. Methods based on nonnegative matrix factorization are also used, but they have high computational complexity. Algorithms based on label propagation lack this disadvantage. Methods based on affiliation model are perspective. This model claims that communities define the structure of a graph.

Algorithms which use node attributes are considered: ones based on latent Dirichlet allocation, initially used for text clustering, and CODICIL, where edges of node content relevance are added to the original edge set. 6 classes are listed for algorithms for graphs with node attributes: changing egdes' weights, changing vertex distance function, building augmented graph with nodes and attributes, based on stochastic models, partitioning attribute space and others.

Overlapping community detection algorithms which effectively use node attributes are just started to appear. Methods based on partitioning attribute space, latent Dirichlet allocation, stochastic models and nonnegative matrix factorization are considered. The most effective algorithm on real datasets is CESNA. It is based on affiliation model. However, it gives results which are far from ground truth covers. Almost all algorithms don't consider the possibility of node attributes partial absence. So one can expect the improvement of existing methods and appearance on new ones.

This work was supported by RFBR (grant No 16-29-09517 "OFI M")

References

1. Fortunato S.. Community detection in graphs. *Physics Reports*, 2010, vol. 486, no. 3-5, pp. 75–174.
2. Papadopoulos S., Kompatsiaris Y., Vakali A., Spyridonos P. Community detection in Social Media. *Data Mining and Knowledge Discovery*, 2012, vol. 24, no. 3, pp. 515–554.
3. Korshunov A.V. Zadachi i metody opredelenija atributov pol'zovatelej sotsial'nyh setej [Problems and methods of the atribute definition of users of social networks]. *Trudy 15 Vserossijskoj nauchnoj konferentsii "Elektronnye biblioteki: perspektivnye metody i tehnologii, elektronnye kollektsii"*, RCDL'2013. 2013. (in Russian)
4. 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. *CoRR*, 2008, vol. abs/0810.1355.
5. Newman M.E.J., Girvan M. Finding and evaluating community structure in networks. *Physical Review E*, 2004, vol. 69, iss. 2, art. no. 026113 (15 pages). DOI: 10.1103/ PhysRevE.69.026113
6. Schaeffer S.E.. Graph clustering *Computer Science Review*, 2007, vol. 1, no. 1, pp. 27–64.
7. Scott J.. *Social network analysis: A handbook. 2nd ed*. London, SAGE, 2000.
8. Newman M.E.J. Fast algorithm for detecting community structure in networks. *Physical Review E*, 2004, vol. 69, iss. 6, art.no. 066133 (5 pages). DOI: 10.1103/PhysRevE.69.066133
9. Kannan R., Vempala S., Vetta. A. On clusterings: Good, bad and spectral. *Journal of the ACM*, 2004, vol. 51, no. 3, pp. 497-515.
10. Hoory S., Linial N., Wigderson A. Expander graphs and their applications. *Bulletin of American Mathematical Society*, 2006, vol. 43, no. 4, pp. 439-561. DOI: 10.1090/S0273-0979-06-01126-8
11. Rosvall M., Bergstrom C.T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 2008, vol. 105, no. 4, pp. 1118–1123.
12. 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, art. no. 036106 (11 pages).
13. Palla G., Derenyi I., Farkas I., Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 2005, vol. 435, no. 7043, pp. 814–818.
14. Xu X., Yuruk N., Feng Z., Schweiger T.A.J. SCAN: A Structural Clustering Algorithm for Networks. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'07. New York, NY, USA, ACM, 2007, pp. 824–833.
15. Yang J., McAuley J.J., Leskovec J. Community Detection in Networks with Node Attributes. *CoRR*, 2014, vol. abs/1401.7267.

16. Xie J., Kelley S., Szymanski B.K. Overlapping Community Detection in Networks: The State-of-the-art and Comparative Study. *ACM Comput. Surv*, 2013, vol. 45, no. 4, pp. 43:1–43:35.

17. Nicosia V., Mangioni G., Carchiolo V., Malgeri M. Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment*, 2009, vol. 2009, no. 3, art.no. P03024.

18. Lazar A., Abel D., Vicsek T. Modularity measure of networks with overlapping communities. *EPL (Europhysics Letters)*, 2010, vol. 90, no. 1, art. no. 18001. DOI: 10.1209/0295-5075/90/18001

19. Busun N.O., Korshunov A.V. Vyjavlenie peresekajushchihsja soobshchestv v sotsial'nyh setjah [Identifying overlapping communities in social networks] *Doklady Vserossijskoj nauchnoj konferentsii "Analiz izobrazhenij, setej i tekstov"*, AIST'2012. 2012. (in Russian).

20. Ahn Y.-Y.Y., Bagrow J.P., Lehmann S. Link communities reveal multiscale complexity in networks. *Nature*, 2010, vol. 466, no. 7307, pp. 761–764.

21. Lancichinetti A., Fortunato S., Kertesz J. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 2009, vol. 11, no. 3, art.no. 033015.

22. Zhao K., Zhang S., Pan Q. Fuzzy analysis for overlapping community structure of complex network. *Control and Decision Conference*, CCDC. IEEE Computer Society, 2010, pp. 3976–3981.

23. Psorakis I., Roberts S., Ebden M., Sheldon B. Overlapping community detection using Bayesian non-negative matrix factorization. *Physical Review E*, 2011, vol. 83, iss. 6, art. no. 066114 (9 pages). DOI: 10.1103/PhysRevE.83.066114

24. Xie J., Szymanski B. K., Liu X. SLPA: Uncovering Overlapping Communities in Social Networks via a Speaker-Listener Interaction Dynamic Process. *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ICDMW '11. Washington, DC, USA, IEEE Computer Society, 2011, pp. 344–349.

25. Buzun N, Korshunov A., Avanesov V., Filonenko I., Kozlov I., Turdakov, D., Kim, H. EgoLP: Fast and Distributed Community Detection in Billion-Node Social Networks. *2014 IEEE International Conference on Data Mining Workshop*, 2014, pp. 533–540.

26. Yang J., Leskovec J. Overlapping Community Detection at Scale: A Nonnegative Matrix Factorization Approach. *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM'13. New York, NY, USA, ACM, 2013, pp. 587–596.

27. Yang J., Leskovec J. Community-Affiliation Graph Model for Overlapping Network Community Detection. *12th IEEE International Conference on Data Mining*, ICDM-2012. Brussels, Belgium, December 10-13, 2012, pp. 1170–1175.

28. Airoldi E.M., Blei D.M., Fienberg S.E., Xing E.P. Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, 2008, vol. 9, pp. 1981–2014.

29. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 2003, vol. 3, p. 993–1022.

30. Ruan Y., Fuhry D., Parthasarathy S. Efficient Community Detection in Large Networks Using Content and Links. *Proceedings of the 22nd International Conference on World Wide Web*, WWW'13. New York, NY, USA, ACM, 2013, pp. 1089–1098.

31. Chesnokov V.O., Klyucharev P.G. Social Graph Community Differentiated by Node Features with Partly Missing Information. *Science and Education of the Bauman MSTU*, 2015, no. 9, pp. 188–199. DOI: 10.7463/0915.0811704 (in Russian).

32. Bothorel C., Cruz J.D,Magnani M., Micenkova B. Clustering attributed graphs: Models, measures and methods. *Network Science*, 2015, vol. 3, no. 3, pp. 408–444.

33. Neville J., Adler M., Jensen D. Clustering relational data using attribute and link information. *Proceedings of the Text Mining and Link Analysis Workshop, 18th International Joint Conference on Artificial Intelligence*, 2003, pp. 9–15.

34. Zhou Y., Cheng H., Yu J.X. Graph Clustering Based on Structural/Attribute Similarities. *Proc. VLDB Endow*, 2009, vol. 2, no. 1, pp. 718–729.

35. Li R., Wang C., Chang K.C.-C. User Profiling in an Ego Network: Co-profiling Attributes and Relationships. *Proceedings of the 23rd International Conference on World Wide Web*, Seoul, Korea, ACM, 2014, pp. 819–830.

36. Huang X., Cheng H., Yu J.X. Dense Community Detection in Multi-valued Attributed Networks. *Inf. Sci.*, 2015, vol. 314, no. C, pp. 77–99.

37. Gunnemann S., Boden B., Farber I., Seidl T. Efficient Mining of Combined Subspace and Subgraph Clusters in Graphs with Feature Vectors. *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference*, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part I. Springer Berlin Heidelberg, 2013, pp. 261–275.

38. Balasubramanyan R., Cohen W.W. Block-LDA: Jointly Modeling Entity-Annotated Text and Entity-Entity Links. *Handbook of Mixed Membership Models and Their Applications*, 2014, pp. 255–273.

39. Nguyen H. T., Dinh T. N. Unveiling the structure of multi-attributed networks via joint non-negative matrix factorization. *Military Communications Conference*, MILCOM 2015. IEEE, 2015, pp. 1379–1384.

40. Hu L., Chan K.C. C. Fuzzy Clustering in a Complex Network Based on Content Relevance and Link Structures. *IEEE Transactions on Fuzzy Systems*, 2016, vol. 24, no. 2, pp. 456–470.