Сетевое программирование

Лекция 10

Сообщества и кластеры в сетях. Понятие сообщества. Алгоритмы выявления сообществ (Girvan–Newman, модульность)

Усіпов Н.М., PhD · КазНУ имени аль-Фараби

Цель и план лекции

Цель

Изучение понятия сообщества и функции качества разбиения сети, освоение алгоритмов выявления сообществ в NetworkX и понимание применения кластеризации при анализе и проектировании телекоммуникационных сетей.

1. Введение
2. Понятие сообщества
3. Модульность
4. Алгоритмы выявления сообществ
5. Применение кластеризации в телекоммуникациях
6. Спектральные и статистические методы
7. Модификации алгоритмов и сравнение разбиений
8. Сегментирование сети на основе сообществ

1 Введение

- Многие реальные сети имеют выраженную модульную структуру: узлы образуют группы, внутри которых связи расположены плотно, а между группами встречаются редко.
- Разделение сети на сообщества отличается от разбиения на компоненты связности: сообщества обычно связаны между собой.

2 Понятие сообщества

Интуитивно сообщество представляет собой подмножество вершин, для которого плотность внутренних связей заметно выше плотности связей с внешней частью сети.

Разбиение

Каждый узел принадлежит ровно одному сообществу. Это наиболее распространённая постановка задачи.

Перекрывающиеся сообщества

Узел может входить сразу в несколько групп. Такая ситуация типична для граничных маршрутизаторов, связывающих различные сегменты.

Иерархические сообщества

Крупные группы состоят из более мелких: например, регион включает в себя города, а город – районы.

3 Модульность

- Для количественной оценки качества разбиения Ньюман и Гирван предложили функцию модульности:

$$Q = (1 / 2m) \Sigma [A_{ij} - k_i k_j / 2m] \delta(c_i, c_j),$$

- где m – число рёбер, A_{ij} – элемент матрицы смежности, k_i – степень вершины i, а $\delta(c_i, c_j)$ равно единице, если вершины i и j находятся в одном сообществе, и нулю в противном случае.
- Таким образом, модульность сравнивает плотность внутри сообществ с ожидаемой при случайном соединении.

4 Алгоритмы выявления сообществ

Алгоритм Гирвана–Ньюмана (2002)

Основан на том, что рёбра между сообществами имеют высокую центральность по посредничеству.

Жадная оптимизация модульности

Изначально каждая вершина образует отдельное сообщество. На каждом шаге объединяется пара, дающая наибольшее увеличение Q. Функция: `greedy_modularity_communities`.

Алгоритм Лувена (2008)

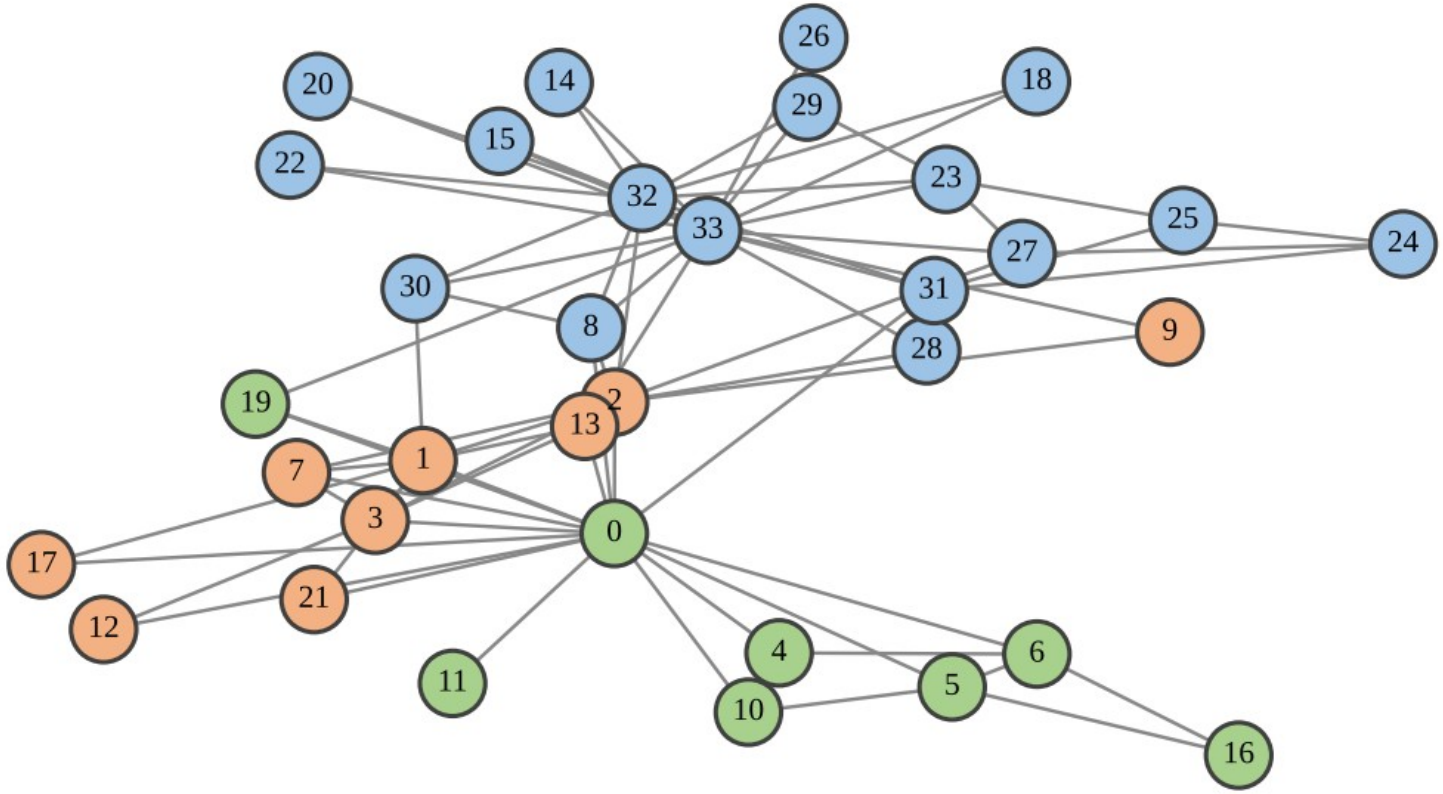
Состоит из двух повторяющихся фаз: локального перемещения вершин и агрегации сообществ в новый граф.

Распространение меток

Каждая вершина принимает метку, наиболее распространённую среди соседей. Метод очень быстрый, однако результат может меняться от запуска к запуску.

Цветом вершин обозначена принадлежность к сообществу.

4 Алгоритмы выявления сообществ (продолжение)



Сообщества в сети «клуб каратэ», выявленные жадной оптимизацией модульности

4 Выявление сообществ в сети «клуб каратэ»

Листинг 10.1.

```
import networkx as nx
from itertools import islice

G = nx.karate_club_graph()
nx.set_edge_attributes(G, 1, "weight")    # сеть без весов
Q = nx.community.modularity

gn = nx.community.girvan_newman(G)
two = next(gn)
print("Girvan-Newman: групп", len(two),
      "размеры", sorted(map(len, two)), "Q =", round(Q(G, two), 3))

best = max(islice(nx.community.girvan_newman(G), 8),
           key=lambda c: Q(G, c))
print("GN, лучшее Q:", len(best), "групп, Q =", round(Q(G, best), 3))

gr = nx.community.greedy_modularity_communities(G)
print("Жадный:", len(gr), "групп, Q =", round(Q(G, gr), 3))

lv = nx.community.louvain_communities(G, seed=1)
print("Лувен:", len(lv), "групп, Q =", round(Q(G, lv), 3))
```

4 Результат: Выявление сообществ в сети «клуб каратэ»

```
Girvan-Newman: групп 2 размеры [15, 19] Q = 0.36
GN, лучшее Q: 5 групп, Q = 0.401
Жадный: 3 групп, Q = 0.381
Лувен: 4 групп, Q = 0.419
```

Обратите внимание на то, что первое разбиение алгоритма Гирвана–Ньюмана не обязательно является наилучшим по модульности: для выбора числа сообществ необходимо просмотреть ряд уровней иерархии.

5 Применение кластеризации в телекоммуникациях

Иерархическая маршрутизация

Деление сети на области уменьшает объём служебной информации. Например, в OSPF маршрутизаторы одной области обмениваются полной топологической базой, а между областями передаются сводные сведения.

Планирование и сегментирование

Группы тесно взаимодействующих устройств размещают в одном сегменте, что уменьшает трафик между сегментами.

Размещение контроллеров и кэшей

В каждое сообщество помещается отдельный сервер для приближения сервисов к пользователям.

Анализ трафика и безопасность

Появление необычных групп взаимодействующих узлов может указывать на сбой или активность вредоносной программы.

Для проверки алгоритма удобно использовать искусственные сети с известной структурой.

5 Применение кластеризации в телекоммуникациях (продолжение)

Кластеризация в сенсорных сетях

Узлы группируются вокруг головного узла для экономии энергии при передаче данных.

Для проверки алгоритма удобно использовать искусственные сети с известной структурой.

5 Проверка на сети с заданной структурой

Листинг 10.2.

```
G = nx.planted_partition_graph(4, 25, 0.5, 0.02, seed=3)
true = G.graph["partition"]
found = nx.community.louvain_communities(G, seed=3)

print("Найдено групп:", len(found))
print("Q заданного разбиения:", round(Q(G, true), 3))
print("Q найденного разбиения:", round(Q(G, found), 3))

same = sum(set(a) == set(b) for a in true for b in found)
print("Полностью совпавших групп:", same)
```

```
Найдено групп: 4
Q заданного разбиения: 0.628
Q найденного разбиения: 0.628
Полностью совпавших групп: 4
```

6 Спектральные и статистические методы

- Кроме алгоритмов, описанных выше, существует ряд других подходов к выявлению сообществ.
- Спектральная кластеризация применяет собственные векторы матрицы Лапласа. Вектор Фидлера, соответствующий второму наименьшему собственному значению, позволяет приближённо разбить граф на две части с малым числом связей между ними.
- Стохастическая блочная модель (SBM) описывает вероятность появления ребра как функцию принадлежности концов к группам.
- Подход Infomap (Росвалл и Бергстром, 2008) исходит из теории информации: случайное блуждание по сети кодируется минимальным числом бит, а хорошим разбиением считается то, что обеспечивает самое короткое описание.

7 Модификации алгоритмов и сравнение разбиений

- Алгоритм Лейдена (Траг и др., 2019) устраняет недостаток алгоритма Лувена, при котором могут образовываться плохо связанные или даже разорванные сообщества.

$$Q_\gamma = (1/2m)\,\Sigma\,[\,A_{ij} - \gamma\,k_i k_j/2m\,]\,\delta(c_i, c_j).$$

- При увеличении γ образуются более мелкие сообщества. Это позволяет получить разбиение нужного масштаба и частично преодолевает проблему предела разрешения.
- Сравнение различных разбиений выполняется с помощью нормализованной взаимной информации (NMI) и скорректированного индекса Рэнда (ARI).

7 Модификации алгоритмов и сравнение разбиений (продолжение)

- Модульность может иметь множество почти одинаковых по значению разбиений, заметно отличающихся друг от друга.

8 Сегментирование сети на основе сообществ

- При разбиении корпоративной или операторской сети на области алгоритм выявления сообществ следует дополнить инженерными условиями.

9 Разбор примера

1. Рассмотрим сеть из двух треугольников {1, 2, 3} и {4, 5, 6}, соединённых одним ребром 3–4.
2. Всего m = 7 рёбер.
3. Для вычисления модульности удобно использовать формулу $Q = \Sigma[e_c/m - (d_c/2m)^2]$, где e_c – число рёбер внутри сообщества, а d_c – сумма степеней его вершин.
4. Для каждого треугольника $e_c = 3$, а $d_c = 2 + 2 + 3 = 7$.
5. Вклад равен $3/7 - (7/14)^2 = 0{,}4286 - 0{,}25 = 0{,}1786$.
6. Итоговое значение $Q = 2 \cdot 0{,}1786 \approx 0{,}357$.
7. Если объединить все вершины в одно сообщество, то $Q = 7/7 - 1 = 0$.

9 Разбор примера (продолжение)

8 Следовательно, разделение на два треугольника значительно лучше.

10 Вопросы для обсуждения на занятии

Граница групп

Какие элементы сети находятся на границе сообществ и почему они важны для надёжности?

Ограничения

Какие технические требования должны дополнять алгоритм при разделении сети на области OSPF?

Проверка

Как убедиться, что найденные сообщества не являются артефактом алгоритма?

Контрольные вопросы

1. Что такое сообщество в сети? Чем оно отличается от компоненты связности?
2. Какие виды структуры сообществ существуют?
3. Что показывает модульность и как интерпретировать её значение?
4. В чём заключается предел разрешения модульности?
5. Опишите принцип работы алгоритма Гирвана–Ньюмана.
6. Как работает алгоритм Лувена и в чём его преимущества?
7. Какие задачи телекоммуникаций решаются методами выявления сообществ?
8. Как можно проверить качество работы алгоритма выявления сообществ?

Список литературы

1. Girvan M., Newman M.E.J. Community structure in social and biological networks // Proceedings of the National Academy of Sciences. – 2002. – Vol. 99, No. 12. – P. 7821–7826.
2. Newman M.E.J., Girvan M. Finding and evaluating community structure in networks // Physical Review E. – 2004. – Vol. 69. – Art. 026113.
3. Clauset A., Newman M.E.J., Moore C. Finding community structure in very large networks // Physical Review E. – 2004. – Vol. 70. – Art. 066111.
4. Blondel V.D., Guillaume J.-L., Lambiotte R., Lefebvre E. Fast unfolding of communities in large networks // Journal of Statistical Mechanics: Theory and Experiment. – 2008. – Art. P10008.
5. 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. – Art. 036106.
6. Fortunato S., Barthélemy M. Resolution limit in community detection // Proceedings of the National Academy of Sciences. – 2007. – Vol. 104, No. 1. – P. 36–41.
7. Zachary W.W. An information flow model for conflict and fission in small groups // Journal of Anthropological Research. – 1977. – Vol. 33, No. 4. – P. 452–473.
8. Easley D., Kleinberg J. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. – Cambridge: Cambridge University Press, 2010.
9. NetworkX Documentation [Электронный ресурс]. – URL: https://networkx.org/documentation/stable/