# Адаптивный рандомизированный алгоритм выделения сообществ в графах

Тимофей Проданов

---

Слайд 1

Добрый день, я Тимофей Проданов и я буду рассказывать об адаптивном выделении сообществ в графах.

Последние двадцать лет развивается изучение сложных сетей, то есть графов с неправильной, сложной структурой. Сложные сети были с успехом применены в различных областях, например в биоинформатике и социологии. Такие сети, построенные по реальным системам, часто разбиваются на группы тесно связанных узлов графа, сообщества. Способность находить и анализировать подобные группы узлов даёт большие возможности в изучении реальных систем, например, сообщества в интернете представляют распространённые темы.

---

Слайд 2
Блондель

Винсент Блондель в 2008 году разбил на сообщества два миллиона абонентов бельгийской телефонной компании. На рисунке размер круга отвечает за размер сообщества, а цвет за основной язык — французский или голландский. Заметно, как зелёные и красные сообщества распределелись по разным сторонам сети, а в центральных сообществах цвета оказались смешанными.

---

Слайд 3
Выделение сообществ

Было показано, что сложные сети, построенные по реальным системам часто имеют характеристики, усложняющие анализ классической теорией графов или статистический анализ. В 2004 году Марк Ньюман представил целевую функцию модулярность, показывающую качество разбиения графа на сообщества. Модулярность удобна в использовании, так как подсчёт выигрыша от объединения двух сообществ занимает одну операцию, а объединение двух сообществ — от количества соседей одного из сообществ. Таким образом, задачу выделения сообществ рассматривают как задачу максимизации модулярности.

---

Слайд 4
Алгоритмы выделения сообществ

Эффективными оказались рандомизированные алгоритмы выделения сообществ в графах. В 2010 году Овельгённе и Гейер-Шульц предложили рандомизированный жадный алгоритм выделения сообществ с параметром $k$. На каждой итерации рассматривается $k$ случайных сообществ, у каждого сообщества исследуются соседи, и затем соединяется лучшая пара.

В 2012 году те же учёные победили на конкурсе DIMACS со схемой кластеризации основных группа графа. Сначала $s$ начальных алгоритмов выделяют сообщества, а те узлы, которые начальные алгоритмы назначили в разные сообщества, распределяет по сообществам финальный алгоритм. В итеративной схеме узлы, относительно которых начальные алгоритмы не сошлись во мнении, снова распределяют по сообществам начальные алгоритмы.

В качестве начальных и финального алгоритмов можно использовать рандомизированный жадный алгоритм с разными параметрами.

---

Слайд 5
*SPSA*. Постановка задачи

Однако не существует набора параметров, при хоторых эти алгоритмы хорошо выделяют сообщества в каждом графе, и остаётся открытым вопрос об версиях алгоритмов, которые были бы работоспособны на большем количестве задач.

Для создания таких модификаций используется стохастический градиентный спуск. Алгоритм *SPSA* предполагает разбиение улучшаемого алгоритма на два $n$ шагов. В течении нечётного шага улучшаемый алгоритм использует один параметр, а во время чётного — другой, после чего подбираются следующие два значения параметра.

В работе рассматривается применение *SPSA* к двум рандомизированным алгоритмам выделения сообществ для получения модификаций, способных хорошо работать на большем количестве входных графов.

Слайд 6
Адаптивный рандомизированный жадный алгоритм

Предлагается адаптиный рандомизированный жадный алгоритм или $ARG$, почти идентичный рандомизированному жадному алгоритму с параметром $k$. Однако действие $ARG$ разбито на шаги длиной в $\sigma$ итераций, после окончания шага параметр $k$ меняется.

Для получения следующих двух параметров $k$ некоторая текущая оценка параметра возмущается в обе стороны. Следующая оценка будет ближе к тому возмущению, которое дало лучшие результаты. [Оценка результата зависит от $k$ и от медианы прироста модулярности за $\sigma$ шагов.]

---

Слайд 7
Параметры $ARG$

В отличии от $RG$, имевшего один параметр, $ARG$ имеет пять параметров, однако от чувствительности и количества итераций в шаге результат зависит слабо, а размер возмущения и начальная оценка имеет значения, дающие хороший результат на всех графах. Существует параметр, отвечающий за значимость времени, при его увеличении время работы уменьшается, однако при уменьшении модулярности.

---

Слайд 8
Сравнение $RG$ и $ARG$

Для сравнения качества рандомизированного жадного алгоритма и его адаптивной версии сопоставлялась медианная модулярность рзбиений. На тепловой карте изображены модулярности разбиений графов с конкурса DIMACS. Левее пунктирной линии указаны $RG$ с разными параметрами, правее — $ARG$ с разными параметрами. Чем более зелёный цвет — тем разбиение лучше, и чем краснее — тем хуже. Видно, что $RG$ даёт плохие результаты чаще, чем $ARG$. [$RG_{10}$]

---

Слайд 9
Применение $ARG$ в $CGGC$

Первым этапом $CGGC$ $s$ начальных алгоритмов выделяют сообщества. Если начальный алгоритм даёт плохие разбиения, то и результат всей схемы будет плохим. Однако по значению модулярности невозможно сказать, относительно хорошее это разбиение или относительно плохое.

Поэтому имеет смысл использовать $ARG$ вместо $RG$ в качестве начального алгоритма, ведь нестабильные начальные алгоритмы портят финальное разбиение.

На таблице изображены результаты $CGGC$, в верхней строке указаны начальные алгоритмы, во второй — финальные, а слева — три тестовых графа. Золотым цветом отмечены лучшие результаты, серебрянным — вторые по модулярности, а красным — очень плохие. Видно, что в последних трёх столбцах с начальным алгоритмом $ARG$ показаны более хорошие значения, и ни разу не появляются очень плохие.

---

Слайд 10
$ACGGC$

Представляется адаптивная схема кластеризации основных групп графа, схожая с неадаптивной схемой, но на первом этапе в качестве начальных алгоритмов используется $RG$ с подстраивающимися параметрами.

Затем в создании промежуточного разбиения участвуют не все начальные разбиения, но только некоторое количество лучших.

---

Слайд 11
Параметры $ACGGC$

$ACGGC$ имеет несколько параметров: размер возмущения, чувствительность, начальную оценку, долю хороших параметров и количество шагов. Однако можно подобрать набор параметров, достаточно хорошо работающих на всех тестовых графах.

---

Слайд 12
Снижение времени работы

Предложено два механизма снижения времени работы. Первый из них заключается в возможности увеличения параметра значимости времени, уменьшая время и модулярность. Второй механизм устанавливает ограничение на максимальную оценку параметра $k$, этот вариант более стабилен и часто даже улучшает результаты при уменьшении времени работы. Снизу изображены две тепловые карты для второго механизма, слева — модулярность при уменьшении максимальной оценки слева направо, а справа — время при уменьшении параметра.

---

Слайд 13
Сравнение $ACGGC$ и $CGGC$

На таблице слева указаны тестовые графы с конкурса DIMACS, от 34 до полутора миллионов узлов. Два первых столбца представляют результаты $ACGGC$ с разными параметрами, а следующие три — результаты $CGGC$ с разными параметрами. Золотым отмечены лучшие результаты, серебрянным — вторые по модулярности и красным — очень плохие результаты.

$ACGGC$ даёт лучше и более стабильные значения.

---

Слайд 14
Сравнение $ACGGCi$ и $CGGCi$

Как и $CGGC$, $ACGGC$ можно итерировать. На таблице изображены результаты итерационных схем, слева направо дважды $ACGGCi$ с разными парамететрами, $CGGCi$ и комбинированный из двух алгоритмов вариант в последнем столбце.