ВЫДЕЛЕНИЕ СООБЩЕСТВ В ГРАФАХ В ЗАДАЧАХ КОМПЬЮТЕРНОЙ КРИМИНАЛИСТИКИ

Куликов Е.К., студент кафедры системного программирования СПбГУ,
ek@belkasoft.com

Аннотация

В статье проводится обзор некоторых существующих алгоритмов выделения сообществ в графах, а также рассматривается конкретное практическое приложение этой задачи: предлагается алгоритм выделения сообществ для графов, возникающих в результате проведения компьютерно-технических экспертиз в рамках следственных мероприятий. Этот алгоритм показал достаточно точные результаты на большом наборе тестовых данных и приемлемую скорость работы, что позволяет использовать его в современном продукте цифровой криминалистики.

Введение

Выделение сообществ в графах (англ. community detection) [1], то есть разбиение множества вершин графа на подмножества (возможно, пересекающиеся), так что в каждом сообществе число рёбер, исходящих в другие его вершины значительно превышает количество тех, что связывают вершины этого сообщества с остальными вершинами графа, является важным инструментом анализа социальных сетей, применяется также при визуализации больших графов, в социологии.

Исследования в данной области начались в середине прошлого века и продолжаются по сей день, достигнув максимальной активности в середине 2000-х годов. Тогда эти разработки не получили широкого применения в криминалистике, так как социальные сети вроде Facebook, число пользователей которой исчисляется миллионами, только зарождались, а смартфоны с мобильным интернетом и вовсе не существовали, поэтому объём электронного общения был сравнительно небольшим, пользователи в большинстве своём имели лишь один адрес электронной почты или ICQ. Сейчас же объём общения посредством интернета значительно вырос, поэтому исследования в области выделения сообществ в графах стали представлять для компьютерной криминалистики серьёзный интерес.

Особо важную роль выделение сообществ может сыграть при анализе графов, отражающих взаимодействия членов террористических организаций, зачастую имеющих сложную структуру и большое число участников, в том числе географически удалённых друг от друга. Выявление групп связанных участников могло бы помочь предсказать планируемые противоправные деяния, в том числе теракты, и предотвратить их, а также задержать виновных в уже произошедшем преступлении, не задержанных или уничтоженных непосредственно на месте преступления.

Большой вклад в развитие этой теории внесли такие учёные, как M. Newman, A. Lancichietti, S. Fortunato. К сожалению, число публикаций отечественных специалистов по данной задаче крайне мало.

Постановка задачи и обзор

Целью данной работы является построение алгоритма выделения сообществ в графах, возникающих при проведении компьютерно-технической экспертизы электронных устройств, показывающего достаточно точные результаты и приемлемую скорость работы.

Ни один из существующих на сегодняшний день инструментов цифровой криминалистики (FTK, Nuix, EnCase, и др.) интеллектуальный анализ социальных графов не поддерживает, кроме получения списка смежных вершин по нажатию на данную (Nuix). Таким образом, возможность выделения сообществ в построенных графах взаимодействий между лицами, причастными к совершению преступления каким-либо способом, остутствует.

Тем не менее, существует достаточно большое число научных работ теоретического и практического характера, посвящённых задаче выделения сообществ в графах. В 2009 году итальянский учёный S. Fortunato опубликовал работу под названием “Community detection in graphs”, в которой было собрано подавляющее большинство известных на тот момент алгоритмов выделения сообществ.

Исследования, в которых проводится сравнительный анализ различных методов разбиения графа на сообщества, существуют (работы A. Lancichinetti и S. Fortunato [2], G.Orman et al. [3], К.А. Славнова [4]), однако их число невелико, и они имеют существенные недостатки.

Так, модельные графы, используемые во втором и третьем из вышеперечисленых исследований (Girwan-Newman Benchmark, l-partition model, Lancichinetti model), являются невзвешенными, в то время как в задачах компьютерной криминалистики крайне важно учитывать,

насколько тесным было взаимодействие, то есть вес ребра, вычисленный при построении графа по специальному алгоритму. Ни одна из статей не анализирует алгоритм Prat-Perez et al. [5] и некоторые другие, разработанные после 2011 года. Кроме того, авторы статей приходят к несколько различным выводам: А. Лансичинетти в своей работе утверждает, что Infomap - один из лучших методов, тогда как по мнению К.А. Славнова, качество результатов этого алгоритма весьма невысокое. Тем не менее, все три исследования сходятся на том, что использование Louvain метода [6] позволяет получать достаточно качественные результаты.

Результаты исследования

В рамках проводимого исследования мною были реализованы несколько алгоритмов выделения сообществ (Algorithm of Girwan and Newman, Algorithm by Radicchi et al., Lovain method, Baumes and Goldberg overlapping method, Prat-Perez et al. Algorithm, Markov Cluster Algorithm) [1], использующих различные подходы к решению задачи и критерии качества разбиения, допускающих или не допускающих перекрывающиеся сообщества (англ. overlapping communities). Алгоритмы были протестированы на наборе социальных графов, полученных в результате проведения компьютерно-технических экспертиз.

Louvain метод на криминалистических данных также показал достаточно точные результаты и приемлимую скорость работы. Попытки использования остальных алгоритмов привели к неудаче либо из-за крайне низкой их производительности (Algorithm of Girwan and Newman), либо невысокого качества получаемого разбиения. Так, многие алгоритмы (особенно Markov Cluster Algorithm и Algorithm by Radicchi et al.) часто выдают в качестве результата большое количество достаточно маленьких сообществ при отсутствии крупных. Такое разбиение не позволяет делать серьёзные выводы об исследуемом графе и взаимодействиях реальных лиц, которые он отражает.

Однако классический Louvain method также имеет недостаток. Достаточно часто к правильному выделенному сообществу достаточно большого размера алгоритм добавляет также несколько вершин, фактически к этому сообществу не относящихся, но смежных с одной из его вершин.

Решить эту проблему позволяет следующий подход: начальным набором сообществ в алгоритме следует считать не множество вершин графа (то есть каждая вершина — отдельное сообщество, как это предполагается в классическом описании алгоритма), а набор сообществ,

построенный алгоритму Prat-Perez et al. Этот метод наряду с Louvain обладает достаточно высокой скоростью анализа графа, и даже допускает распараллеливание некоторых его этапов, поэтому использование сразу двух алгоритмов не оказывает серьёзного влияния на производительность, однако позволяет получать (как показало тестирование на графах, полученных при проведении компьютерно-технических экспертиз) более точные разбиения графа на сообщества, нежели при использовании исключительно Louvain method.

Заключение

В статье предлагается алгоритм выделения сообществ в графах, показавший достаточно точные результаты и приемлемую для использования в современном продукте цифровой криминалистики производительность, опробованный на наборе графов, возникавших при проведении реальных компьютерно-технических экспертиз.

Литература

1. Fortunato S. Community detection in graphs, PhysicsReports, 2010.
2. Lancichinetti A. и Fortunato S. Community detection algorithms: A comparative analysis, Physical review, 2009.
3. Orman G., Labatut V. и Cherifi H. Qualitative comparison of community detection algorithms // International Conference on Digital Information and Communication Technology and its Applications – с.265—279 – 2011.
4. Славнов К. А. Анализ социальных графов — 2015. http://www.machinelearning.ru/wiki/images/6/60/2015_417_SlavnovKA.pdf [дата просмотра 28.03.2016].
5. Prat-Perez A., Domingues-Sal D. и Larriba-Pei L. High quality, scalable and parallel community detection for large real graphs // Proceedings of the 23rd international conference on World wide web – с.225—236 – 2014.
6. Blondel V., Guillaume J., Lambiotte R. Fast unfolding of communities in large networks // An IOP and SISSA journal – 2008.