CORECLUSTER: A Degeneracy Based Graph Clustering Framework

Christos Giatsidis
École Polytechnique
giatsidis@lix.polytechnique.fr

Dimitrios M. Thilikos
CNRS, LIRMM and UoA
sedthilk@thilikos.info

Fragkiskos D. Malliaros
École Polytechnique
fmalliaros@lix.polytechnique.fr

Michalis Vazirgiannis
École Polytechnique and AUEB
mvazirg@lix.polytechnique.fr

Abstract

Graph clustering or community detection constitutes an important task for investigating the internal structure of graphs, with a plethora of applications in several domains. Traditional tools for graph clustering, such as spectral methods, typically suffer from high time and space complexity. In this article, we present CORECLUSTER, an efficient graph clustering framework based on the concept of graph degeneracy, that can be used along with any known graph clustering algorithm. Our approach capitalizes on processing the graph in a hierarchical manner provided by its core expansion sequence, an ordered partition of the graph into different levels according to the k-core decomposition. Such a partition provides a way to process the graph in an incremental manner that preserves its clustering structure, while making the execution of the chosen clustering algorithm much faster due to the smaller size of the graph’s partitions onto which the algorithm operates.

Introduction

Detecting clusters or communities in graphs constitutes a cornerstone problem with many applications in several disciplines. Characteristic application domains include social and information network analysis, biological networks, recommender systems and image segmentation. Due to its importance and multidisciplinary nature, the problem of graph clustering has received great attention from the research community and numerous algorithms have been proposed (see (Fortunato 2010) for a survey in the area).

Spectral clustering methods (e.g., (Ng, Jordan, and Weiss 2001)) impose a high cost of computing resources both in time and space regardless of the data on which it is going to be applied (Fortunato 2010). Other well-known approaches for community detection are the ones based on modularity optimization (Newman and Girvan 2004; Clauset, Newman, and Moore 2004), stochastic flow simulation (Satuluri and Parthasarathy 2009) and local partitioning methods (Fortunato 2010). In any case, scalability is still a major challenge in the graph clustering task, especially nowadays with the significant increase of the graphs’ size.

Typically, there are two main methodologies for scaling up a graph clustering method: (i) algorithm-oriented and (ii) data-oriented. The first one considers the algorithm of interest and appropriately optimizes – whenever is possible – the “parts” of the algorithm responsible for scalability issues. Prominent examples here are the fast modularity optimization method (Clauset, Newman, and Moore 2004) and the scalable flow-based Markov clustering algorithm (Satuluri and Parthasarathy 2009). The second and widely used methodology is to rely on sampling/sparsification techniques. In this case, the size of the graph onto which the algorithm will operate is reduced, by disregarding nodes/edges. However, in this approach possibly useful structural information of the graph (i.e., nodes/edges) is ignored.

In this paper, we propose CORECLUSTER, a graph clustering framework that capitalizes on the notion of graph degeneracy – also known as k-core decomposition (Seidman 1983). The main idea behind our approach is to combine any known graph clustering algorithm with an easy-to-compute, clustering-preserving hierarchical representation of the graph – as produced by the k-core decomposition – towards a scalable graph clustering tool. The k-core of a graph is a maximal size subgraph where each node has at least k neighbors in the subgraph (we say that k is the rank of such a core). The maximum k for which a graph contains a k-core is known as its degeneracy. We refer to this core as “the densest core”. Intuitively, the k-core of such a graph is located in its “densest territories”. Based on this idea, we show that the densest cores of a graph are roughly maintaining its clustering structure and thus constitute good starting points (seed subgraphs) for computing it. Given the fact that the size of the densest core of a graph is orders of magnitude smaller than that of the original graph, we apply a clustering algorithm starting from its densest core and then, on the resulting structure, we incrementally cluster the rest of the nodes in the lower rank cores in decreasing order – following the hierarchy produced by the k-core decomposition.

The main contributions of this paper are the following:

- Clustering Framework: We introduce CORECLUSTER, a scalable degeneracy-based graph clustering framework, that can be used along with any known graph clustering algorithm. We show how CORECLUSTER utilizes the k-core decomposition of a graph in order to (i) select seed subgraphs for starting the clustering process and (ii) expand the already formed clusters or create new ones.

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Scalability and Accuracy Analysis: We discuss analyti-
cally the ability of CoreCluster to scale-up, describ-
ing its expected running time. We also justify why the
$k$-core structure captures the clustering properties of a
graph, thus being able to indicate good seed subgraphs
for a clustering algorithm.

Experiments: We perform an extensive experimental
evaluation regarding the efficiency and accuracy of
CoreCluster, both on synthetic and real-world graphs.
The experimental results show that the time complexity
is improved by 3-4 orders of magnitude (compared to a
baseline algorithm), especially for large graphs. More-
ever, for graphs with inherent community structure, the
quality of the results is maintained or even is improved.

Related Work

Graph clustering. The problem of community detection
and graph clustering has been extensively studied from
several points of view. Some well-known approaches in-
clude spectral clustering (e.g., (Ng, Jordan, and Weiss 2001;
Shi and Malik 2000; White and Smyth 2005)), modularity
optimization (e.g., (Newman 2004; Clauset, Newman,
and Moore 2004)), multilevel graph partitioning (e.g., Metis
(Karypis and Kumar 1998)), flow-based methods (Satuluri
and Parthasarathy 2009), hierarchical methods (Newman
and Girvan 2004) and many more. A very informative and
comprehensive review over the different approaches can
be found in (Fortunato 2010). Also, the authors of (Lan-
cichinetti, Fortunato, and Radicchi 2008) have conducted
a comparative analysis on the performance of some of the
most recent algorithms, in artificial data produced by their
parametrized generator of benchmark graphs. In our work,
we use the same graph generator as in (Lancichinetti, Fortu-
nato, and Radicchi 2008) to evaluate our framework.

Scaling-up graph clustering. The efficiency of graph clus-
tering can be improved in various ways. Two well-known
approaches are the ones of sampling and sparsification. In
the case of spectral clustering, sampling-based approaches
include the Nyström method (Kumar, Mohri, and Talwalkar
2009) and randomized SVD algorithms (Drineas et al.
2004). Concerning graph sampling, the goal is to produce
a graph of smaller size (nodes and edges), preserving a set
of desired graph properties (e.g., degree distribution, clus-
tering coefficient) (Leskovec and Faloutsos 2006). The work
by Maiya and Berger-Wolf (Maiya and Berger-Wolf 2010),
presents a method – based on the notion of expansion prop-
eties – to sample a subgraph that preserves the community
structure, i.e., contains representative nodes of the commu-
nities. Then, the community membership of the nodes that
do not belong to the sample can be expressed as an inference
problem. Unlike the aforementioned methods that sample
both nodes and edges, the graph sparsification algorithm pre-
sented in (Satuluri, Parthasarathy, and Ruan 2011) reduces
only the number of edges (focusing on inter-community
edges) in order improve the running time of a clustering
algorithm. In contrast to the above methods, our approach
keeps the structure of the graph intact, without excluding
any structural information from the clustering process. $k$-core decomposition. Seidman (Seidman 1983) first ap-
plicated the $k$-core decomposition to study the cohesion of
social networks. Since then, it has been applied in several
graph-related tasks, such as graph visualization (Alvarez-
Hamelin et al. 2005; Zhang and Parthasarathy 2012), as an
edge ordering criterion for graph coarsening (Abou-Rjeili
and Karypis 2006), and in the decomposition of massive
graphs (Cheng et al. 2011).

Preliminaries

Given a graph $G$, we denote by $V(G)$ and $E(G)$ the sets of
its vertices and edges respectively. Given a set $S \subseteq V(G)$,
we denote by $G[S]$ the subgraph of $G$ that is obtained if we
remove from it all vertices that do not belong in $S$. We also
use $n$ and $m$ for the number of the vertices and edges of $G$,
i.e., $n = |V(G)|$ and $m = |E(G)|$. The neighborhood of a
vertex $v \in V(G)$ in $G$ is denoted by $N_G(v)$ and contains
all vertices of $G$ that are adjacent to $v$. The degree $\deg_G(v)$
of a vertex $v$ in $G$ is equal to $|N_G(v)|$. The minimum degree
of $G$, denoted by $\delta(G)$, is the minimum degree of the
vertices in $G$, i.e., $\delta(G) = \min\{\deg_G(v) \mid v \in V(G)\}$. Given
a non-negative integer $k$, we define its $k$-core, denoted by
$\core_k(G)$ as the maximum size subgraph of $G$ with
minimum degree $k$ and we say that $k$ is the rank of this core.
The degeneracy, denoted by $\deg^*(G)$ of a graph $G$ is the max-
imum $k$ for which $G$ contains a non-empty core. In other
words, $\deg^*(G) = \max\{\delta(H) \mid H \subseteq G\}$.

Intuitively, the dense cores of a graph, i.e., those whose
ranks are close to $\deg^*(G)$, may serve as a good seeds of start-
ing any clustering algorithm as we expect them to respect
the clustering structure of the original graph. Later at this
paper, we make this statement more precise, providing the
necessary theoretical justification.

For graph $G$ and $\deg^*(G) = k$, we define its core expansion
sequence as the sequence of vertex sets $\{V_k, V_{k-1}, \ldots, V_0\}$
that is recursively defined as follows: $V_k = V(\core_k(G))$
and $V_i = V(\core_i(G)) \backslash V_{i+1}$, $i = k-1, \ldots, 0$. We refer
to the sets of a core expansion sequence as layers, with set
$V_i$ being its $i$-th layer (see Fig. 1).

Detecting the $i$-core of a graph $G$ is easy (Batagelj and
Zaversnik 2003): just remove vertices of degree less than
$i$ until this is not possible any more. It can be easily seen
that the computation of $k$-core decomposition can be done
in $O(k \cdot n)$ steps. As real world graphs have typically small
degeneracy, this makes the computation of the core expan-
sion sequence an easy computational task.

Proposed Method

CoreCluster capitalizes on the concept of degeneracy to
improve the efficiency of graph clustering. The main idea
behind our approach is that the $k$-core decomposition pre-
serves the clustering structure of a graph and therefore
the “best” $k$-core subgraph can be used as good starting point for
a clustering method. Furthermore, the decomposition pro-
vides an hierarchical organization of the nodes in the graph,
that can “guide” the clustering process.
that uses clustering of the graph. Suppose we have an algorithm that takes as input a graph $G$ and outputs a partition of $V(G)$ into a number of sets that form a clustering of $G$. As in this section we are mainly focused on the general aspects of our method, we do not further specify the attributes of such an algorithm and we, abstractly, name it $\text{Cluster}$. We also assume that it runs in $O(n^3)$ steps as, in our experiments, the $\text{Cluster}$ algorithm is the spectral algorithm of (Ng, Jordan, and Weiss 2001). Our aim is to define a procedure — called $\text{CoreCluster}$ — that uses $\text{Cluster}$ and accelerates the algorithm without any significant loss in its accuracy.

In simple terms, the $\text{CoreCluster}$ framework applies the $\text{Cluster}$ algorithm at the highest $k$-core of the graph and then it iterates from the highest to the lowest core trying to apply the following logic: assign with a simple criterion all the nodes that can be assigned to the existing clusters and apply the $\text{Cluster}$ algorithm to the remaining nodes (in order to create new clusters).

**Procedure** $\text{CoreCluster}(G)$.

**Input:** A graph $G$.

**Output:** A partition of $V(G)$ into clusters.

1. $k := 2^0(G)$.
2. $q := 0$.
3. Let $V_0, \ldots, V_6$ be the core expansion sequence of $G$.
4. For $i = 0, \ldots, k$, let $G_i$ be the $i$-core of $G$.
5. Let $S_k = V_6$.
6. Let $A_k = \{C_1^k, \ldots, C_{n_k}^k\} = \text{Cluster}(G[S_k])$.
7. For $i = k-1$ to $0$ do
8. \[ S_i := \text{Select}(G_i, A_k \cup \ldots \cup A_{i+1}, V_i). \]
9. let $A_i = \{C_1^i, \ldots, C_{n_i}^i\} = \text{Cluster}(G[S_i])$.
10. Return $A_0 \cup \ldots \cup A^0$.

Initially, $\text{CoreCluster}$ performs $k$-core decomposition to obtain the core expansion sequence of the graph. Then, algorithm $\text{Cluster}$ is applied to the $k$-core subgraph, creating the first clusters. The procedure $\text{Select}$ takes as input the, so far, created clusters, i.e., the sets in $F_{i+1} = A_k \cup \ldots \cup A_{i+1}$ and the $i$-layer $V_i$ and tries to assign each of the vertices of $V_i$ in some cluster in $A_k \cup \ldots \cup A_{i+1}$. After this update, the procedure $\text{Select}$ returns the unassigned vertices. The choice of the selection procedure considers the way the vertices of $V_i$ are adjacent with the vertices of the clusters in $A_k \cup \ldots \cup A_{i+1}$. This selection can be done with several heuristic approaches, and next we describe such a procedure.

We stress that $\text{CoreCluster}$ can be essentially seen as a “meta-algorithmic procedure” in the sense that it can be applied to any clustering algorithm. The discussion that follows in the next parts of the paper argues that this indeed can improve the clustering argument in time without any significant expected loss in its performance.

**Selection procedure**

In this section we describe the selection procedure ($\text{Select(\_)}$) in Line 8 of $\text{CoreCluster}$. The procedure takes as input the so far created clustering $F_{i+1} = A_k \cup \ldots \cup A_{i+1}$ and the vertex set $V_i$. We describe below how this procedure assigns some of the vertices of $V_i$ to the clusters in $F_{i+1}$ and outputs the remaining ones.

First of all we call a pair $(G, F, V)$ a candidate triple, if $G$ is a graph, and $F \cup \{V\}$ is a partition of $V(G)$. Given a candidate triple $(G, F, V)$, we define the following property on the vertices of $V$:

\[
\mathbf{P}^\alpha,\beta(v) = \exists C \in F : \frac{|N_G(v) \cap V(C)|}{|N_G(v)|} \geq \alpha
\]

and $|N_G(v)| \geq \beta$,

where $\alpha > 0.5$ and $\beta$ is a positive integer. Notice that, as $\alpha > 0.5$, the truth of $\mathbf{P}^\alpha,\beta(v)$ can be certified by a unique set $C$ in $F$. We call such a set the certificate of $v$.

**Procedure** $\text{Select}(G, F, V)$.

**Input:** A candidate triple $(G, F, V)$.

**Output:** A subset $S$ of $V$ and a partition $F'$ of $V(F) \cup \{V \setminus S\}$.

1. while $\mathbf{P}^\alpha,\beta(v)$ is true for some $v \in V$,
2. set $F' := (F \setminus \{C\}) \cup \{C \cup \{v\}\}$ where
3. $C$ is the certificate of $v$
4. and set $V' := V \setminus \{v\}$.
5. set $V'' := N_G(V(F))$ and $V^2 := V \setminus (V' \cup F')$.
6. if $V^2$ is either empty or an independent set of $G$,
7. then
10. return $\emptyset$.
11. else return $V' \cup V''$.

Before we present the $\text{assign}$ routine we need some definitions. Given a candidate triple $(G, F, V)$ and a vertex $v \in V$ we define $\text{span}(v) = \max\{|N_G(v) \cap V(C)| : C \in F\}$. We also define $\text{argspan}(v)$ as a minimum size $C \in F$ with the property that $|N_G(v) \cap V(C)| = \text{span}(v)$.

**Procedure** $\text{assign}(G, F, V, S)$.

**Input:** A candidate triple $(G, F, V)$ and a subset $S$ of $V$.

**Output:** A partition $F'$.

1. while $S \neq \emptyset$,
2. let $l = \max\{\text{span}(v) : v \in S\}$
3. let $L = \{v \in S : \text{span}(v) = l\}$
4. for every $v \in L$,
5. set $C' = C \cup \{v\}$ where $C = \text{argspan}(v)$
6. set $S := S \setminus \{v\}$
7. set $F' := (F \setminus \{C\}) \cup \{C'\}$
8. return $F'$. 46
The selection procedure first (lines 1–4) tries to assign vertices of \( V \) to clusters of \( \mathcal{F} \) using the criterion of the property \( P^{\alpha, \beta} \) that assigns a vertex to a cluster only if the vast majority of its neighbors belong in this cluster. The quantification of this “vast majority” criterion is done by the constants \( \alpha \) and \( \beta \). The vertices that cannot be assigned are partitioned into two groups: \( V^1 \) contains those that have neighbors in vertices that are already classified in the clusters of \( \mathcal{F} \) and \( V^2 \) contains the rest. As the vertices in \( V^2 \) have no neighbors in the clusters, they have at least \( k \) neighbors out of them, it is most likely that they may not enter to any existing cluster in the future, unless, possibly, they are completely disjoint. If this is not the case, a further (milder) classification is attempted by the assign procedure that first classifies the vertices in \( V^1 \) in the existing clusters and then we do the same for the vertices in \( V^2 \). We stress that this last selection has been useful in our experiments in cores of low rank (where many independent vertices may appear). The procedure assign is a heuristic that classifies each vertex to the cluster that has the majority of its neighbors.

**Expected execution time**

The speed up of the algorithm is based on the fact that CORECLUSTER\((G)\) now runs in \( k + 1 \) disjoint subgraphs of \( G \) instead from \( G \) itself. As the \( i \)-th selection phase requires \( O(|V(G,i)|^2) \) steps, we conclude that the running time of CORECLUSTER\((G)\) is bounded by

\[
\sum_{i=k,...,0} O(|S_i|^3) \leq \sum_{i=k,...,0} O(|V_i|^3) \leq O(k \cdot n_{\max}^3),
\]

where \( n_{\max} = \max\{|V_k|, \ldots, |V_0|\} \). In the above bound, the first equality holds only in the extremal case where no selection occurs during the selection phases. Clearly, the general bound in Eq. (1) is the best possible when \( |V_k|, \ldots, |V_0| \) tend to be equally distributed (which would accelerate the running time by a factor of \( k^2 \)). According to the first inequality of Eq. (1) the running time of the algorithm is proportional to \((k + 1) \cdot n_{\max}^3\), where \( n_{\max} = \max\{|S_k|, \ldots, |S_0|\} \). Let \( \rho_G = \max\{|V(G)| / |S_i| | i = 0, \ldots, k\} \) and \( \mu_G = \max\{|V(G)| / |V_i| | i = 0, \ldots, k\} \) and observe that \( \rho_G \geq \mu_G \). Notice that the discrepancy between \( \rho \) and \( \mu \) is a measure of the acceleration of the algorithm because of the selection phases.

Concluding, the acceleration of CORECLUSTER is upper bounded by

\[
\sum_{i=k,...,0} O(|S_i|^3) = O\left(\frac{k}{\rho^3} \cdot n^3\right).
\]

This estimation is purely theoretical and its purpose is to expose the general complexity contribution of our algorithmic machinery. In practice, the acceleration can be much better and this also depends on the heuristics that we apply for the selection phase.

**Quality of the CORECLUSTER framework**

The intuition behind our framework is that the core expansion sequence \( V_k, V_{k-1}, \ldots, V_0 \) gives a good sense of direction on how to perform clustering in an incremental way. After that, the procedure considers \( V_{k-1} \) as the remaining vertices of the \((k - 1)\)-core \( G_{k-1} \), and tries to assign them one by one to the already existing clusters \( C^1_k, \ldots, C^{\rho_k}_k \). The vertices for which such an assignment is not possible, form the set \( S_{k-1} \) and the Cluster is now applied on \( G[S_{k-1}] \). As the algorithm continues, the existing clusters grow up and the vertices for which this is not possible, are grouped to new clusters. The fact that this procedure approximates satisfactorily the result of the application of Cluster to the whole graph is justified by the observation that the early \( i \)-cores (i.e., \( i \)-cores where \( i \) is close to \( k \)) are already dense, and therefore sufficiently coherent, to provide a good starting clustering that will expand well because of the selection criterion. In fact, the subgraphs obtained by the \( k \)-core decomposition, provide an \((1/2)\)-approximation algorithm for the DENSEST-SUBGRAPH problem (Andersen and Chellapilla 2009).

**Theoretical justification.** We claim that the decomposition identifies subgraphs that progressively correspond to the most central regions and connected parts of the graph. Here we show that nodes with high clustering coefficient (Watts and Strogatz 1998) in \( G \), are more likely to survive at the highest \( k \)-core subgraph by the pruning (\( k \)-core decomposition) procedure. Our claim is based on the following theorem:

**Theorem 1 (Gleich and Seshadhri 2012)** Let \( G \) be a graph with heavy-tailed degree distribution, and let \( C_G \) be the (global) clustering coefficient of \( G \). Then, there exists a \( k \)-core in \( G \) for \( k \geq C_G d_{\max}^2 / 2 \), where \( \epsilon < 1 \) is a constant such that most edges are incident to a node with degree at least \( d_{\max}^2 \) (typically \( \epsilon = 2/3 \)), where \( d_{\max} \) is the maximum degree of the nodes.

The above theorem implies that graphs with heavy-tailed degree distribution and high global clustering coefficient \( C_G \), have large degeneracy. Next we present our claim for the relationship between the local clustering coefficient \( C_v \), \( \forall v \in V(G) \) and the \( k \)-core subgraph justifying the selection of the \( k \)-core as good seed subgraph in the clustering procedure.

**Claim 1** Let \( G \) be a graph with heavy-tailed degree distribution. The contribution of each node \( v \in V(G) \) to the \( k \)-core decomposition of the graph is proportional to the local clustering coefficient \( C_v \).

**Proof Sketch.** The global clustering coefficient \( C_G \) of the entire graph is given by the average of the local clustering coefficients \( C_v \), \( \forall v \in V(G) \), i.e., \( C_G = \frac{1}{n} \sum_C C_v \), where \( n = V(G) \). Then, from Theorem 1 we have that:

\[
k \geq C_G \frac{d_{\max}^2}{2} = \left( \frac{1}{n} \sum_C C_v \right) \frac{d_{\max}^2}{2} = \left( \frac{1}{n} \frac{d_{\max}^2}{2} \right) \sum_C C_v
\]

\[
\Rightarrow k \geq \gamma \sum_C C_v,
\]

where parameter \( \gamma \) captures global characteristics of the graph (that depend on the total number of nodes and the
maximum degree). Therefore, nodes with high clustering coefficient (in the original graph) are more likely to be found in the best k-core \((k = \delta^*(G))\) subgraph, since they tend to be more robust to the degeneracy process. Thus, the k-core subgraph can be used as a good starting point (seed subgraph) for the clustering task. Additionally, we have experimentally validated the above claim (see (Giatsidis et al. 2014)).

**Experimental Evaluation**

Here we present the experimental results of our framework on both the amelioration of the execution time and the quality of the clustering results. Based on parameter space exploration, the values for the parameters of the Select procedure are chosen to be \(\alpha = 0.8\) and \(\beta = 5\) (this choice appears to work optimally in our experiments).

**Spectral algorithm**  As the baseline and the basis for the CORECLUSTER framework (algorithm Cluster), we use the Ng-Jordan-Weiss spectral clustering algorithm as it is described in (Ng, Jordan, and Weiss 2001). The basic idea of this algorithm is to keep the top \(k\) eigenvectors of the normalized adjacency matrix and perform \(k\)-means clustering on the rows of the matrix composed from these eigenvectors. In our variation we are using \(k\)-means++ (Arthur and Vassilvitskii 2007) for its advantage of performing better seeding during the initialization process and, since we desire to have an automatic choice for \(k\), we define \(k\) by the "eigengap" as it is suggested in (Polito and Perona 2001).

**Datasets description**

While real networks are the objective, actual datasets lack ground truth which leaves only evaluation metrics of the quality of clustering as an option and not direct comparison. On the other hand, artificial networks offer ground truth and a large variety of properties that can be parameterized to produce different "types" of networks. The evaluation of our framework is conducted on both real and artificial networks in order to have complete and decisive results.

**Artificial Networks.** We exploit the graph generator proposed in (Lancichinetti, Fortunato, and Radicchi 2008) to produce graphs with ground truth clustering structure. This graph generator provides a wide range of input parameters. We used the parameters in Table 1 and tuned them for various combinations in order to get a wide range of graphs with different features. Thus, the testing of our approach is credible as it is evaluated in essentially hundreds of graphs with different properties and quality of clustering structure. The parameters used are: \(N\) is the size of the graph, \(max_d\) is the maximum node degree, \(min_d\) is the minimum node degree (this is the most important parameter as it is the one differentiating the overall density of the graph) and \(\mu\) is the mixing parameter representing the overlapping between clusters, i.e., each node shares a fraction \(1 - \mu\) of its links with the other nodes of its community and a fraction \(\mu\) with the other nodes of the network.

**Real Networks.** We also perform evaluations to a subset of the Facebook dataset (Traud, Mucha, and Porter 2011). This is a collection of friendship networks of Facebook from 2005, for 100 US Universities (i.e., 100 individual networks). The evaluations were not performed to the full extent of this dataset as hardware limitation did not allow us to evaluate, with spectral clustering, networks with more than 13\(K\) nodes (the CORECLUSTER framework could handle much larger networks). About half of the networks from this dataset were used for the final evaluation.

**Time performance**

Figures 2 (a) and 2 (b) depict the specific execution times for the artificial graph D1 and Facebook respectively.

<table>
<thead>
<tr>
<th>Graph Size</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>2000</td>
<td>10</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>3000</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 1: Parameters’ values for the artificial graphs.

From Figure 2, it is evident that the gain in execution time using the proposed framework is very significant (i.e., at least three orders of magnitude for graph sizes above 3000 nodes) and increases polynomially with the graph size. Additionally, as we can see in Figure 3, the execution time of our framework increases rather linearly with the graph size – thus showing good scaling features.

**Clustering quality evaluation and setup**

For each of the graphs at hand we run i. as baseline approach the Ng-Jordan-Weiss (Ng, Jordan, and Weiss 2001) spectral graph clustering algorithm and ii. our CORECLUSTER framework on the datasets presented earlier. Following, we describe the methods and metrics for evaluating the results on artificial and real networks.

**Artificial networks:** We measure the quality of the clustering results in terms of the widely used Normalized Mutual Information (Manning, Raghavan, and Schütze 2008) (NMI). In
For reasonable values of the mixing parameters, the overlap of the clusters is low. Of course the counterargument here is that for larger values of the mixing parameter, the overlap of the clusters is such that it basically prevents the definition of a clustering structure – and therefore perhaps it is meaningless to search for clusters in these cases.

We have to stress that in all cases, the execution time of our algorithm, especially for large graphs, is 3-4 orders of magnitude smaller that those of the plain clustering algorithms – achieving essentially the same or even better quality for reasonable values of the mixing parameters.

**Facebook:** The networks of this dataset lack ground truth, and for this reason we choose to evaluate the results with the evaluation criterion of conductance. Given a graph \( G \) and a cut \( (S, \overline{S}) \), conductance is defined as

\[
\phi(S) = \frac{\sum_{i \in S, j \notin S} A_{ij}}{\min(\alpha(S), \alpha(\overline{S}))},
\]

where \( A_{ij} \) are the entries in the adjacency matrix \( A \) of \( G \) and \( \alpha(S) = \sum_{i \in S} \sum_{j \in G} A_{ij} \). Informally, conductance measures (for a cluster) the ratio of internal to external connectivity. It has been used widely to examine clustering quality (e.g., (Leskovec, Lang, and Mahoney 2010)) and has a simple and intuitive definition. In Fig. 4 (b), we can see the comparison of conductance values versus different sizes of detected communities by the two methods. Conductance has values in the range \((0, 1)\) with lower values indicating better clustering quality.

For better presentation (in Fig. 4 (b)) we have aggregated the detected cluster sizes (in terms of number of nodes) into bins of 500 (e.g., 0–500, 501–1000, etc.) and have provided the average conductance for each bin. This plot essentially provides the comparison of average clustering quality between the baseline and CORE CLUSTER for different cluster sizes. Before commenting on the comparison, it is important to note that – for both methods – clusters with less than 10 nodes were excluded as they were trivial with regards to the clustering criteria for large scale graphs. Moreover, we have evaluated CORE CLUSTER to a larger subset of Facebook including networks that we could not evaluate with the baseline spectral, due to limitations of hardware memory. Consequently, we have results of clusters up to \( 8K \) nodes (from networks of up to \( 13K \) of nodes) for the baseline and results of clusters up to \( 16K \) (from networks of up to \( 23K \) of nodes) for CORE CLUSTER.

Moving on to the comparison, we can see that CORE CLUSTER displays better clustering quality than the baseline, with the exception of the first bin. The difference is negligible and only slightly surpassed by the baseline’s conductance value. For the last two bins of the baseline, we should note that there was only one cluster found for each, with the one having zero conductance consisting of the entire network (i.e., the whole graph was found as one cluster). In fairness, we could consider an “in between” value but it would be still worse than the corresponding conductance of CORE CLUSTER. Overall, we see that CORE CLUSTER displays a quite low conductance regardless of cluster size, indicating better clustering results in much faster time.

**Conclusions**

In this paper, we articulate an effort for optimizing the efficiency of graph clustering, capitalizing on the intuition that the extreme \( k \)-core of a graph preserves the clustering structure of the original graph, while it is much faster to execute clustering on this degenerate graph due to its much smaller size. Our main contribution is the CORE CLUSTER framework, that initiates clustering on the highest rank core of the graph and then incrementally clusters the graph’s nodes in the subsequent lower rank cores. Moreover we described analytically why this framework scales-up the clustering process and we showed experimentally on a multitude of graph data that the framework decreases the execution time while...
maintaining (or even improving) the clustering quality.

Acknowledgments

The authors would like to thank the anonymous reviewers for the constructive comments. Christos Giatsidis and Michalis Vazirgiannis were partially supported by the DIGITEO Chair grant LEVETONE in France. Fragkiskos D. Malliaros is a recipient of the Google Europe Fellowship in Graph Mining, and this research is supported in part by this Google Fellowship. Dimitrios M. Thilikos was co-financed by the E.U. (European Social Fund - ESF) and Greek national funds through the Operational Program “Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF) - Research Funding Program: “Thales. Investing in knowledge society through the European Social Fund”.

References


