

Generalized K-means for Metric Space Clustering Using PageRank

Mustafa Hajj¹, Eyad Said², Robert Todd²

¹Ohio State University & Columbus Ohio, USA

²Mount Mercy University, Cedar Rapid, Iowa

Abstract

We utilize the PageRank vector to generalize the k -means clustering algorithm to directed and undirected graphs. We demonstrate that PageRank and other centrality measures can be used in our setting to robustly compute centrality of nodes in a given graph. Furthermore, we show how our method can be generalized to metric spaces and apply it to other domains such as point clouds and triangulated meshes.

1. Introduction

The k -means algorithm is one of the most well-studied and popular point cloud clustering algorithms [Jai10]. The generic version of k -means algorithm takes as input a point cloud X and the number of clusters k and returns a partition of X into k subsets or clusters. Due to its popularity, the k -means algorithm has been studied extensively in the clustering literature and many variations of it have been suggested [Che03; Che06; DGK04] including kernel versions [DGK04]. See also [Jai10] and the references therein.

This article addresses the graph clustering problem by leveraging centrality measures defined on the nodes of a given graph. More specifically, we utilize the *PageRank* vector [BP98] and view it as a centrality measure on a graph to generalize the k -means clustering algorithm to graphs. The algorithm introduced here is applicable to directed and undirected graphs.

Graph clustering algorithms have a vast literature. The reader is referred to other sources for more details [Sch07]. Multiple graph clustering algorithms have been suggested over the past few decades including spectral-based methods [DGK04], minimal spanning tree-based methods [Stu03] and clique-based methods [ESB99]. See also [Sch07] for other clustering methods on graphs.

The k -means algorithm is related to the Voronoi diagrams which can be defined in the context of graphs [ZDKG12]. Voronoi diagrams on graphs have been utilized for finding meaningful clusters in biological networks for example. [ZDKG12].

Several other works have found ways to apply the k -means algorithm (Lloyd's algorithm) to graphs of various flavors [BKZ15; WB11]. As noted in [Erw00], we computing the Voronoi cells for directed and undirected graph is not particularly challenging despite the fact that we cannot define a metric for a general directed graph (and thus applying the k -means algorithm). The main challenge in this context is the definition of the centroid of a

cluster. The problem of defining the centroid was addressed with the Karcher/Fréchet mean [GK73; Pen06] when considering point clouds in Riemannian manifolds. Other heuristics to define the center in a metric space setting are also possible [DF85]. K -medoids clustering, as a discrete version of k -means [RMJ07], is another method of addressing the problem of centroids in a discrete setting. These definitions are however very expensive to compute, in particular in the context of the k -means algorithm, and are not applicable directly to general directed graphs. The theory in [CLCL16] provides definitions that allow one to apply a k -means-like algorithm for computing clusters in any metric space without defining centroids. However, other definitions are required. In [GT10] the personalized PageRank vector [BCG10] was utilized to give a graph visualization algorithm for a k -means based clusters of the graph nodes. However, there algorithm is only applicable to graphs and does not generalize to metric spaces.

Our novel contribution is to use nodes with high centrality measure in lieu of the centroid of a cluster in Lloyd's algorithm. While any centrality measure can be used, we focus on PageRank centrality. The algorithm we propose here is completely independent of any embedding of the graph into a metric space and is based only on the graph's connectivity structure. Nevertheless, we can apply this algorithm to graphs that are in a metric space such as a 3D-mesh or the neighborhood graph of point clouds in some Euclidean space.

Our algorithm has several main advantages. First, the PageRank vector can be defined for directed and undirected graphs. Second, that the PageRank vector was designed to be computed on massive graphs provides additional speed. Third, the algorithm we give here can be easily generalized to metric spaces making it applicable to other domains. Finally, the simplicity will be evident when we present the main algorithm.

2. PageRank and Other Centrality Measures

The PageRank function [BP98] defined on the nodes of a graph can be viewed as centrality measure. For a directed graph $G(V, E)$, the PageRank function $PR : V \rightarrow \mathbb{R}$ is defined for every vertex $v \in V$ by $PR(v) = \frac{(1-\alpha)}{|V|} + \alpha \sum_{u \in out(v)} \frac{PR(u)}{|out(u)|}$, where $out(v)$ is the set of nodes connected to v by out edges leaving v ; $0 < \alpha < 1$ is the *damping factor*, typically set at 0.85. When the graph is undirected, a different version of PageRank function is used [Gro12]. The PageRank vector can be computed efficiently by the power method [HF18]. Intuitively, a high PageRank value at a given node v usually means that v is connected to many other nodes, which also have high PageRank scores. From this perspective, PageRank can be viewed as a measure of centrality for the nodes of the graph. See Figure 1 and observe that more central nodes in the graph example tend to have higher PageRank values (indicated with nodes with the red color).

While PageRank provides us with a good and fast measure of centrality for the nodes of the graph, other centrality measure can be utilized in Algorithm 1. In fact, PageRank is a special case of a more general family of centrality measure called eigenvector centrality [Bon87] and these functions can also be used for this purpose.

Other centrality measures can also be utilized for our purpose. This includes harmonic centrality [ML00; Roc09], information centrality [BF05], closeness centrality [Fre79], and VoteRank [ZCDZ16] among many other measures. Figure 1 shows a few example of centrality measures visualized on a graph.

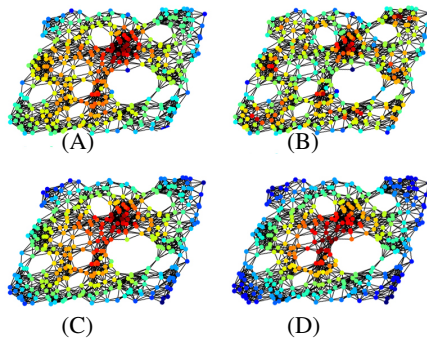


Figure 1: Various centrality measures on graphs. (A) Information centrality, (B) PageRank. (C) Harmonic centrality. (D) Closeness centrality.

3. The Main Algorithm

Just like the traditional k -means algorithm, the algorithm we suggest here have three stages : *the initialization stage, the assignment stage and the update stage*. We discuss these steps in details here. The termination of the algorithm can be chosen to be after a number of iterations or when the Voronoi diagrams do not update for a few consecutive iterations. A summary of the algorithm is given in Algorithm 1. For the remainder of this section we assume that we are given a graph $G = (V, E)$, a method to compute the metric d on G and integer $k > 0$ representing the number of desired clusters. We will not specify if the graph is directed or undirected as all methods introduced here are applicable to both types of graphs.

3.1. The initialization stage

Just like the traditional k -means clustering algorithm, the algorithm we introduce here needs initial choice of the centroids. For our purpose here, we simply choose k nodes from the graph G uniformly at random. We will leave better initialization methods for future work.

3.2. The assignment stage

The assignment stage starts with a list of k centroids of the graph G . We denote these nodes by c_1, \dots, c_k . The metric d along with the nodes c_1, \dots, c_k induce a partition on V called the *Voronoi diagram of the graph* [Erw00].

We recall quickly the definition of a Voronoi diagram on general metric spaces. Let (X, d) be a metric space and let $C \subset X$ be subset of X , called the *subset of centroids*. The *Voronoi cell* at point $c \in C$, denoted by $VC(c)$ is defined to be the set of all points $y \in X$ that are closer to c than to any other point in C . The collection of subsets $VC(c)$ for all c in C is by definition the Voronoi diagram, denoted by $VD(C)$ of the metric space X with respect to the subset C .

In the context of graphs, or other domains such as meshes and point clouds, any choice of a metric d can be deployed and the Voronoi diagram can be computed using optimized algorithms depending on the domain of interest. Our version on graphs uses the algorithm given in [Erw00]. Finally, note that Voronoi diagrams can be computed for directed and undirected graphs. We refer the reader also to [Erw00] for details. In Algorithm 1, at the end of the assignment stage the algorithm returns the Voronoi diagram $VD(\{c_1, \dots, c_k\})$ of the centroids c_1, \dots, c_k , which as we mentioned earlier consists of the sets $V_i := VC(c_i) \subset V$ for each centroid c_i .

3.3. The update stage

The update stage assumes that we are given a partition of the node set : $V_1 \dots V_k$. We use this partition to compute the subgraphs $G_i = (V_i, E_i)$ where $E_i = \{(u, v) \in E | u, v \in V_i\}$ for $1 \leq i \leq k$. We then compute the PageRank $PR_i : V_i \rightarrow \mathbb{R}$ for each G_i . The centroid of each graph G_i is updated simply by $c_i := \operatorname{argmax}_{v \in V_i} (PR_i(v))$. In the rare case when the argmax function returns multiple centroids with the PageRank value, we choose one of these points arbitrarily. Notice the computation of the centroid with this method is reliant of the computation of the PageRank of the subgraphs. The PageRank vector can be computed very efficiently. See [SMPU13] for a $\mathcal{O}(\sqrt{\log(n)}/\epsilon)$ distributed algorithm where n is the number of nodes in the graph and ϵ is fixed constant.

3.4. Distance computation

In the computation of the Voronoi diagram one usually needs to compute the metric d on G . It is important to notice that while the metric d on G is needed for this computation, one usually does not need to compute the entire distance matrix on G .

In our experiments on graphs and triangulated meshes we utilized Dijkstra's algorithm [Dij59] for the metric d . There are multiple methods to speed the distance computations on a graph

Algorithm 1: PageRank-based k -means clustering algorithm on graphs.

Input: Graph $G(V, E)$, number of clusters k .
Output: A partition of the node set V into k subsets.
Initialize the set C by choosing k nodes from V
while While termination criterion has not been met **do**
 for c_i in C **do**
 | Compute $V_i = VC(c_i)$
 end
 for V_i in $VD(C)$ **do**
 | Compute PageRank PR_i on the subgraph (V_i, E_i)
 | $c_i := \text{argmax}_{v \in V_i} (PR_i(v))$
 end
end

[PBCG09]. The heat method for computing geodesics introduced in [CWW13] can also be utilized for fast metric computation on almost all domains that appear in practice. Other metrics that depend on the graph Laplacian can also be utilized. This includes spectral type distances such as commute-time distance, discrete biharmonic distance, and diffusion distance.

4. Extension of the Main Algorithm to Metric Spaces

Algorithm 1 can be easily extended to metric spaces. Indeed, we notice that the initialization and assignment stages that compute the Voronoi cells can be defined for general metric spaces. It is in the update stage where PageRank is utilized. Given that PageRank is just a centrality measure, algorithm 1 generalizes to metric spaces provided the computation of the PageRank function is replaced by an appropriate centrality measure. There are multiple centrality measures that satisfy this criterion e.g. the harmonic centrality and the closeness centrality.

5. Results

To validate the results we applied the main algorithm on several datasets.

We applied our method to some of the graphs available in the NetworkX library [HSS08]. Figure 2 shows the application of Algorithm 1 on various graph examples.

The method that we proposed here is also applicable to point clouds from multiple perspectives. One way to do that is to compute a *neighborhood graph* of the point cloud and then apply algorithm 1 on the graph to obtain a partition of the point cloud. We quickly recall the definition of a neighborhood graph. Let $S \subset \mathbb{R}^n$ be a point cloud with a distance function d_S defined on S . Let $\epsilon > 0$ be a positive number. The neighborhood graph is an undirected graph $\Gamma_{d_S, \epsilon}(S)$, where $\Gamma_{d_S, \epsilon}(S) = (S, E(\Gamma_{d_S, \epsilon}))$ and $E(\Gamma_{d_S, \epsilon}) = \{[u, v] \mid d_S(u, v) \leq \epsilon, u, v \in S, u \neq v\}$. For our computation d_S is simply the Euclidean metric. Figure 3 shows the clusters obtained by applying Algorithm 1 on the neighborhood graph of some point cloud examples. Note that the clusters obtained cannot be usually obtained using traditional k -means algorithm on point cloud with the usual Euclidean distance.

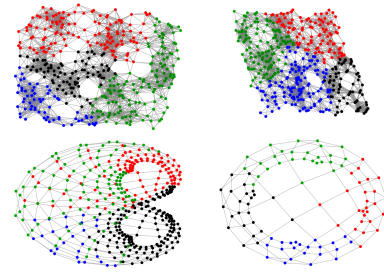


Figure 2: Applying Algorithm 1 on various graph examples. Number of clusters in all examples is 4. The clusters are indicated by the colors of the nodes.

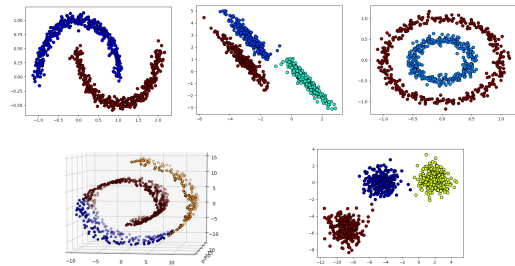


Figure 3: Various example of application of the algorithm on point cloud. For each point the following two steps are applied: (1) The neighborhood graph of the point cloud is calculated. (2) the graph k -means clustering algorithm is applied. The clusters are indicated by the colors of the points.

Finally, mesh segmenting can be considered as a clustering problem. For instance, in [ZL*05] a mesh segmentation is introduced via spectral clustering. See also [Sha08].

In our context, we can view the mesh as a graph and apply Algorithm 1 immediately to this graph. A potentially better approach is to utilize a centrality measures that better describe the geometry of the underlying mesh such as harmonic centrality. Figure 4 show examples of applying Algorithm 1 to triangulated meshes.

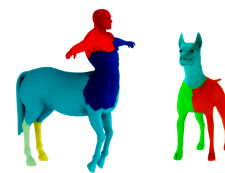


Figure 4: Viewed as metric spaces, we apply algorithm 1 on triangulated meshes. On the left the algorithm is applied with 6 clusters and on the right the algorithm is applied with 3 clusters.

6. Conclusion and Future Work

The method introduced in this paper utilizes centrality measures such as PageRank to generalize the k -means clustering algorithm to graphs. While we explained quickly how our method is applicable to general metric spaces, we have not studied the theoretical

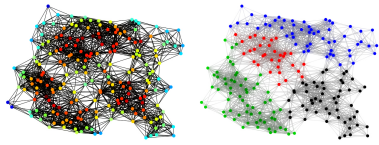


Figure 5: The PageRank function is utilized as a centrality measure in our work. The figure shows the visualization of the PageRank function on the nodes of the graph on the left. On the right we show the application of our algorithm on the same graph with $k = 4$. The clusters are indicated by the colors of the nodes.

properties of the algorithm in this context. Also, it is still not clear which centrality measure yields the best performance under a given metric. We are planning to pursue these directions in future work.

References

- [BCG10] BAHMANI, BAHMAN, CHOWDHURY, ABDUR, and GOEL, ASHISH. "Fast incremental and personalized pagerank". *Proceedings of the VLDB Endowment* 4.3 (2010), 173–184 1.
- [BF05] BRANDES, ULRIC and FLEISCHER, DANIEL. "Centrality measures based on current flow". *Annual symposium on theoretical aspects of computer science*. Springer. 2005, 533–544 2.
- [BKZ15] BÓTA, ANDRÁS, KRÉSZ, MIKLÓS, and ZAVÁLNÍ, BOGDÁN. "Adaptations of the k-means algorithm to community detection in parallel environments". *2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*. IEEE. 2015, 299–302 1.
- [Bon87] BONACICH, PHILLIP. "Power and centrality: A family of measures". *American journal of sociology* 92.5 (1987), 1170–1182 2.
- [BP98] BRIN, SERGEY and PAGE, LAWRENCE. "The anatomy of a large-scale hypertextual web search engine". *Computer networks and ISDN systems* 30.1-7 (1998), 107–117 1, 2.
- [Che03] CHEUNG, YIU-MING. "k-Means: A new generalized k-means clustering algorithm". *Pattern Recognition Letters* 24.15 (2003), 2883–2893 1.
- [Che06] CHEN, KE. "On k-median clustering in high dimensions". *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*. 2006, 1177–1185 1.
- [CLCL16] CHANG, CHENG-SHANG, LIAO, WANJIUN, CHEN, YU-SHENG, and LIU, LI-HENG. "A mathematical theory for clustering in metric spaces". *IEEE Transactions on Network Science and Engineering* 3.1 (2016), 2–16 1.
- [CWW13] CRANE, KEENAN, WEISCHEDEL, CLARISSE, and WARDETZKY, MAX. "Geodesics in heat: A new approach to computing distance based on heat flow". *ACM Transactions on Graphics (TOG)* 32.5 (2013), 1–11 3.
- [DF85] DYER, MARTIN E and FRIEZE, ALAN M. "A simple heuristic for the p-centre problem". *Operations Research Letters* 3.6 (1985), 285–288 1.
- [DGK04] DHILLON, Inderjit S, GUAN, YUQIANG, and KULIS, BRIAN. "Kernel k-means: spectral clustering and normalized cuts". *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004, 551–556 1.
- [Dij59] DIJKSTRA, E. W. "A note on two problems in connexion with graphs". *Numerische Mathematik* 1 (1959), 269–271 2.
- [Erv00] ERWIG, MARTIN. "The graph Voronoi diagram with applications". *Networks: An International Journal* 36.3 (2000), 156–163 1, 2.
- [ESB99] EDACHERY, JUBIN, SEN, ARUNABHA, and BRANDENBURG, FRANZ J. "Graph clustering using distance-k cliques". *International Symposium on Graph Drawing*. Springer. 1999, 98–106 1.
- [Fre79] FREEMAN, L. "Centrality in Networks: I. Conceptual Clarifications. Social Networks". (1979) 2.
- [GK73] GROVE, KARSTEN and KARCHER, HERMANN. "How to conjugate 1-close group actions". *Mathematische Zeitschrift* 132.1 (1973), 11–20 1.
- [Gro12] GROLMUSZ, VINCE. "A note on the pagerank of undirected graphs". *arXiv preprint arXiv:1205.1960* (2012) 2.
- [GT10] GRAHAM, FAN CHUNG and TSIATAS, ALEXANDER. "Finding and visualizing graph clusters using pagerank optimization". *International Workshop on Algorithms and Models for the Web-Graph*. Springer. 2010, 86–97 1.
- [HF18] HOFFMAN, JOE D and FRANKEL, STEVEN. *Numerical methods for engineers and scientists*. CRC press, 2018 2.
- [HSS08] HAGBERG, ARIC, SWART, PIETER, and S CHULT, DANIEL. *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008 3.
- [Jai10] JAIN, ANIL K. "Data clustering: 50 years beyond K-means". *Pattern recognition letters* 31.8 (2010), 651–666 1.
- [ML00] MARCHIORI, MASSIMO and LATORA, VITO. "Harmony in the small-world". *Physica A: Statistical Mechanics and its Applications* 285.3-4 (2000), 539–546 2.
- [PBCG09] POTAMIAS, MICHALIS, BONCHI, FRANCESCO, CASTILLO, CARLOS, and GIONIS, ARISTIDES. "Fast shortest path distance estimation in large networks". *Proceedings of the 18th ACM conference on Information and knowledge management*. 2009, 867–876 3.
- [Pen06] PENNEC, XAVIER. "Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements". *Journal of Mathematical Imaging and Vision* 25.1 (2006), 127 1.
- [RMJ07] RATTIGAN, MATTHEW J, MAIER, MARC, and JENSEN, DAVID. "Graph clustering with network structure indices". *Proceedings of the 24th international conference on Machine learning*. 2007, 783–790 1.
- [Roc09] ROCHAT, YANNICK. *Closeness centrality extended to unconnected graphs: The harmonic centrality index*. Tech. rep. 2009 2.
- [Sch07] SCHAEFFER, SATU ELISA. "Graph clustering". *Computer science review* 1.1 (2007), 27–64 1.
- [Sha08] SHAMIR, ARIEL. "A survey on mesh segmentation techniques". *Computer graphics forum*. Vol. 27. 6. Wiley Online Library. 2008, 1539–1556 3.
- [SMPU13] SARMA, ATISH DAS, MOLLA, ANISUR RAHAMAN, PANDURANGAN, GOPAL, and UPFAL, ELI. "Fast distributed pagerank computation". *International Conference on Distributed Computing and Networking*. Springer. 2013, 11–26 2.
- [Stu03] STUETZLE, WERNER. "Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample". *Journal of classification* 20.1 (2003), 25–47 1.
- [WB11] WITSENBURG, TIJN and BLOCKEEL, HENDRIK. "K-means based approaches to clustering nodes in annotated graphs". *International Symposium on Methodologies for Intelligent Systems*. Springer. 2011, 346–357 1.
- [ZCDZ16] ZHANG, JIAN-XIONG, CHEN, DUAN-BING, DONG, QIANG, and ZHAO, ZHI-DAN. "Identifying a set of influential spreaders in complex networks". *Scientific reports* 6 (2016), 27823 2.
- [ZDKG12] ZIVANIC, MARKO, DAESCU, OVIDIU, KURDIA, ANASTASIA, and GOODMAN, SR. "The Voronoi diagram for graphs and its application in the Sickle Cell Disease research". *Journal of Computational Science* 3.5 (2012), 335–343 1.
- [ZL*05] ZHANG, HAO, LIU, RONG, et al. "Mesh segmentation via recursive and visually salient spectral cuts". *Proc. of vision, modeling, and visualization*. 2005, 429–436 3.