Towards Theoretical Foundations of Clustering: Thesis Highlights

Margareta Ackerman

David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, ON, Canada
mackerma@cs.uwaterloo.ca

ABSTRACT

Clustering is a central unsupervised learning task with a wide variety of applications. However, in spite of its popularity, it lacks a unified theoretical foundation. Recently, there has been work aimed at developing such a theory. We discuss recent advances in clustering theory, including axiomatizing clustering and providing formal guidance for clustering algorithm selection.

This paper presents thesis highlights, summarizing the author's[1] contributions to a formal theory of clustering.

1. INTRODUCTION

Clustering is one of the most widely-used techniques for exploratory data analysis. Given a set of objects, clustering refers to the act of grouping them based on an underlying measure of similarity, aiming to place similar items within the same group. Clustering is used in a wide range of applications, including bioinformatics, computer vision, psychology and marketing, to name a few. Consider, for example, medical research aimed to identify potential risk factors of cancer. Say a survey has been conducted collecting information, such as family history, age, sun exposure, smoking habits, etc. The data can then be clustered, to identify groups (or clusters) of people who are similar based on a combination of these characteristics. Then, if members of a cluster are prone to the same type of cancer, the set of characteristics unifying this cluster are candidate risk factors for the illness.

In spite of hundreds of clustering papers being published every year, its theoretical foundations are meager. Back in 1973, Wright[20] wrote that "while the interest in and application of cluster analysis has been rising rapidly, the abstract nature of the tool is still poorly understood." Three decades later, Kleinberg[13] describes a similar state of affairs, which is still relevant today, "there has been relatively little work aimed at reasoning about clustering independently of any particular algorithm, objective function, or generative data model."

In the last few years there has been a rising interest in developing a theory of clustering. Some interesting new directions of research have emerged and discoveries have been made that both improve our understanding of clustering and have the potential to significantly improve the utility of clustering in practice.

In this paper, we focus on describing our research program and our main contribution towards a unified theory of clustering. We aim towards a better understand of what clustering is, in a general manner, independent of any particular algorithm or generative model. We make progress towards this goal by proposing a consistent set of axioms of clustering quality measures.

While clustering axioms would identify what is common to all clustering functions, concisely formulated properties can also be used to distinguish between different clustering paradigms. Distinguishing algorithms based on their input-output behavior has theoretical interest, but may also have important practical implications.

Faced with a concrete clustering task, a user needs to choose an appropriate clustering algorithm. Currently, such decisions are often made in a very ad hoc, if not completely random, manner. Given the crucial effect of the choice of a clustering algorithm on the resulting clustering, this state of affairs is truly regrettable. We propose a new approach for providing guidance to clustering users by identifying significant properties of clustering functions that, on one hand distinguish between different clustering paradigms, and on the other hand are intended to be relevant to the domain knowledge that a user might have access to. Users could then choose which properties they want an algorithm to satisfy, and determine which algorithms meet their requirements.

In the remainder of this paper, we address four topics in the theory of clustering. In Section 3, we discuss axiomatization of clustering. In Section 4, we elaborate on our main direction of research, which is a property-based classification of clustering paradigms. Section 5 addresses work on notions

[1]Margareta Ackerman is a PhD Candidate, expected to graduate in April, 2012.

of clusterability, and in Section 6 we discuss our characterization of linkage-based clustering, a widely-used clustering paradigm.

2. FORMAL FRAMEWORK

We focus on a basic domain where the input to the clustering function is a finite set of points endowed with a between-points distance (or dissimilarity) function, and the output is a partition of that domain.

A *distance function* is a symmetric function $d : X \times X \to R^+ \cup \{0\}$, such that $d(x, x) = 0$ for all $x \in X$. The data sets that we consider are pairs (X, d), where X is some finite domain set and d is a distance function over X.

A *k-clustering* $C = \{C_1, C_2, \ldots, C_k\}$ of a domain set X is a partition of X into k disjoint subsets of X (so, $\bigcup_i C_i = X$).

A *clustering* of X is a k-clustering of X for some $1 \leq k \leq |X|$.

For a clustering C, let $|C|$ denote the number of clusters in C, and $|C_i|$ denote the number of points in a cluster C_i. For $x, y \in X$ and a clustering C of X, we write $x \sim_C y$ if x and y belong to the same cluster in C, and $x \not\sim_C y$, otherwise.

Definition 1 (clustering function). *A* clustering function *is a function that takes as input a pair (X, d) and a parameter $1 \leq k \leq |X|$ and outputs a k-clustering of the domain X.*

We also consider general clustering functions, whose only input is the domain set.

Definition 2 (General clustering function). *A* general clustering function *is a function that takes as input a pair (X, d) and outputs a clustering of the domain X.*

3. AXIOMATIZING CLUSTERING

This section discusses the work in the following publication: M. Ackerman and S. Ben-David. Measures of Clustering Quality: A Working Set of Axioms for Clustering. NIPS, 2008. [5]

In his highly influential paper, [13], Kleinberg sets up a set of "axioms" aimed to define what a clustering function is. Kleinberg suggests three axioms, each sounding plausible, and shows that these seemingly natural axioms lead to a contradiction - there exists no function that satisfies all three requirements.

Kleinberg's result is often interpreted as stating the impossibility of defining what clustering is, or even of developing a general theory of clustering. We disagree with this view. We show that the impossibility result is, to a large extent, due to the specific formalism used by Kleinberg, rather than being an inherent feature of clustering.

There have been a few other attempts to distill axioms of clustering. Wright[20] did some early work in this direction. In his setting, every domain element is assigned a positive real weight, and its weight may be distributed among multiple clusters. His axioms focus on how clustering functions should handle these weights. More recently, Puzicha et al.[18] consider properties of clustering objective functions, and investigate a class of clustering functions that arises by requiring decomposition into a certain additive form. Also, Meila[15] proposes properties of criteria for comparing clusterings and demonstrates that three such properties cannot be simultaneously satisfied by the same criterion. Of these, Kleinberg's[13] impossibility theorem made the most significant impact on the research community. His elegant and concisely formulated properties put in question the possibility of ever formally defining clustering.

Kleinberg's axioms are intended to capture the meaning of clustering by determining those functions that are worthy of being considered clustering functions and those that are not. He works with general clustering functions whose input is only a domain set, in particular, they do not take the numbers of clusters as part of the input.

The following is Kleinberg's[13] main result.

Theorem 1 (Kleinberg, [13]). *There exists no general clustering function that simultaneously satisfies scale invariance, consistency and richness.*

We include the definitions of the properties, consistency, richness, and scale invariance in Appendix A.

Our contributions: Rather than attempting to define what a clustering function is, and demonstrating a failed attempt, as [13] does, we turn our attention to the closely related issue of evaluating the quality of a given data clustering. We develop a formalism and a consistent axiomatization of that latter notion.

A clustering-quality measure is a function that maps clustering and data set pairs, (X, d) and C, to the set of non-negative real numbers, so that these values reflect how 'good' or 'cogent' that clustering is.

As it turns out, the clustering-quality framework is richer and more flexible than that of clustering functions. In particular, it allows the postulation of axioms that capture the features that Kleinberg's axioms aim to express, without leading to a contradiction. That is, we formulate properties of clustering quality measures that express the principles captured by Kleinberg's axioms, and show that these properties of clustering quality measures are consistent. We demonstrate the relevance and consistency of these axioms by showing that they are satisfied by several natural measures.

4. A FORMAL APPROACH TO CLUSTERING ALGORITHM SELECTION

This section discusses the work in the following publication: M. Ackerman, S. Ben-David, and D. Loker. Towards Property-Based Classification of Clustering Paradigms. NIPS, 2010. [3]

There are many clustering algorithms, and these algorithms often produce different results on the same data. Faced with

Algorithm	outer consistent	inner consistent	local	refinement-confined	order invariant	k-rich	outer rich	inner rich	threshold rich	scale invariant	iso. invariant
Single Linkage	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Average Linkage	✓	*X*	✓	✓	*X*	✓	✓	✓	✓	✓	✓
Complete Linkage	✓	*X*	✓	✓	✓	✓	✓	✓	✓	✓	✓
k-median	✓	*X*	✓	*X*	*X*	✓	✓	✓	✓	✓	✓
k-means	✓	*X*	✓	*X*	*X*	✓	✓	✓	✓	✓	✓
Min sum	✓	✓	✓	*X*	*X*	✓	✓	✓	✓	✓	✓
Ratio cut	*X*	✓	*X*	*X*	*X*	✓	✓	✓	✓	✓	✓
Normalized cut	*X*	*X*	*X*	*X*	*X*	✓	✓	✓	✓	✓	✓

Figure 1: A property-based classification of clustering functions, illustrating what properties are satisfied by some common clustering functions. Properties that distinguish between clustering algorithms can be used to help select an algorithm. Properties that are satisfied by all algorithms are potential axioms.

a concrete clustering task, a user needs to choose an appropriate algorithm. Currently, such decisions are often made in a very ad hoc, if not completely random, manner. Users are aware of the costs involved in employing different clustering algorithms, such as running times, memory requirements, and software purchasing costs. However, there is very little understanding of the differences in the *outcomes* that these algorithms may produce. We propose focusing on that aspect - the input-output properties of different clustering algorithms.

The choice of an appropriate clustering should, of course, be task dependent. A clustering that works well for one task may be unsuitable for another. While some domain knowledge is embedded in the choice of similarity between domain elements (or the embedding of these elements into some Euclidean space), there is still a large variance in the behavior of difference clustering paradigms over a fixed similarity measure.

For some clustering tasks, there is a natural clustering objective function that one may wish to optimize, but very often the task does not readily translate into a corresponding objective function. Often users are merely searching for a meaningful clustering, without a prior preference for any specific objective function. Many (if not most) common clustering paradigms do not optimize any clearly defined objective utility, either because no such objective is defined (like in the case of, say, single linkage clustering) or because optimizing the most relevant objective is computationally infeasible. To overcome computation infeasibility, the algorithms end up carrying out a heuristic whose outcome may be quite different than the actual objective-based optimum (that is the case with the Lloyd method as well as with spectral clustering algorithms). What seems to be missing is a clear understanding of the differences in clustering outputs in terms of intuitive and usable properties.

Some heuristics have been proposed as a means of distinguishing between the output of clustering algorithms on specific data. These approaches require running the algorithms, and then selecting an algorithm based on the outputs that they produce. In particular, clustering quality measures can be used to evaluate the output of clustering algorithms. These measures can be used to select a clustering algorithm by choosing the one that yields the highest quality clustering [19]. However, the result only apply to the specific data sets on which the algorithms were run, and there are no guarantees on the quality of the output of these algorithms on any other data.

Our contributions: We propose a different approach for providing guidance to clustering users by identifying significant properties of clustering functions that, on one hand distinguish between different clustering paradigms, and on the other hand are intended to be relevant to the domain knowledge that a user might have access to. Based on domain expertise users could then choose which properties they want an algorithm to satisfy, and determine which algorithms meet their requirements.

Our vision is that ultimately, there would be a sufficiently rich set of properties that would provide a detailed, property-based, taxonomy of clustering methods, that could, in turn, be used as guidelines for a wide variety of clustering applications.

We propose properties of clustering functions, and present a property-based classification of common clustering functions with respect to the properties that we propose. We have summarized some of our results in Figure 1. The properties shown in Figure 1 appear in Appendix A, and the algorithms in Appendix B.

5. CLUSTERABILITY

This section discusses the work in the following publication: M. Ackerman and S. Ben-David. Clusterability: A Theoretical Study. AISTATS, 2009. [4]

The aim of clustering is to uncover meaningful partitions in data; however, not all data sets have meaningful partitions. Clusterability is a measure of clustered structure in a data

set. So far, clusterability has been used only peripherally and with no theoretical support. We pioneer a theoretical study of clusterability, comparing and analyzing notions of clusterability. We address this issue with generality, aiming for conclusions that apply regardless of any particular clustering algorithm or any specific data generation model.

A variety of notions of clusterability have been proposed in the literature. For instance, Epter et al.[11] measure clusterability as a ratio of the minimum between-cluster distance over the maximum cluster diameter in the clustering that gives the highest such ratio. In an empirical study, Zhang [21] measures clusterability as the average between-cluster distance over the average within-cluster distance.

The separability notion of clusterability, by Ostrovsky, Rabani, Schulman, and Swamy [17], captures how sharp is the drop in the loss function when moving from a $(k-1)$-clustering to a k-clustering. This notion was used in [17] to show that when data is clusterable, then a modified version of the Lloyd method uncovers a clustering with provably near optimal k-means loss.

A line of research by Blum and Balcan follows a similar approach ([7], [8],[6]). They assume the existence of a correct, "target" clustering. They then propose different clusterability criteria, and show that for each, there are some algorithms that uncover the target clustering when the criteria is satisfied.

Our contributions: We survey several notions of clusterability that have been discussed in the literature, as well as propose a new notion of data clusterability. Our comparison of these notions reveals that, although these notions attempt to evaluate the same intuitive property, and all appear to be reasonable, they are pairwise inconsistent. That is, for each pair of notions, there are data sets that are arbitrarily well-clusterable by one of the notions, but poorly clusterable by the other notion. This illustrates the significance of choosing a specific notion of clusterability, as the same results are not necessarily obtainable using a different notion.

We investigate how hard it is to determine the clusterability of a data set. For each notion, we find the hardness of determining whether the clusterability of a data set exceeds a given threshold. In most cases, it turns out that this is an NP-hard problem.

Our analysis of these notions gives rise to an interesting phenomenon. For the notions that we explore, *the more clusterable a data set is, the easier it is (computationally) to find a close-to-optimal clustering of that data.* It has been recently shown that such a property holds with respect to one notion of clusterability, 'k - separability' [17]. We show that this phenomenon holds for other notions as well. In particular, we prove that for well clusterable data, using the clusterability notions we discuss, near-optimal clustering can be *efficiently* computed.

6. CHARACTERIZATION OF LINKAGE-BASED CLUSTERING

This section discusses the work in the following publications:

- M. Ackerman, S. Ben-David, and D. Loker. Characterization of Linkage-based Clustering. COLT, 2010.
- M. Ackerman and S. Ben-David. Discerning Linkage-Based Algorithms Among Hierarchical Clustering Methods. IJCAI, 2011.

Linkage-based clustering is one of the most commonly-used and widely-studied clustering paradigms. They are iterative algorithms that begin by placing each point in a distinct cluster, and then repeatedly merge the closest clusters until a specified number of clusters is formed.

The closest clusters are determined by a linkage function. The linkage functions used by the most common linkage-based algorithms are as follows.

- *Single linkage:* $\min_{a\in A, b\in B} d(a,b)$.
- *Average linkage:* $\frac{\sum_{a\in A, b\in B} d(a,b)}{|A|\cdot|B|}$
- *Complete linkage:* $\max_{a\in A, b\in B} d(a,b)$.

We provide a novel characterization of linkage-based clustering based on natural, concise properties. This is the first characterization of a commonly-used class of clustering algorithms. However, there are a couple of characterizations of the single-linkage algorithm. Single-linkage is a linkage-based algorithm where the distance between a pair of clusters is measured by the length of the smallest edge between the clusters (see Appendix B). In 1975, Jardine and Sibson [12] characterized single linkage among hierarchical algorithms. More recently, Bosagh Zadeh and Ben-David [9] characterize single-linkage for the k-stopping criteria within our framework of clustering functions.

Our contributions: We provide a surprisingly simple set of properties that, on one hand is satisfied by all the algorithm in that family, while on the other hand, no algorithm outside that family satisfies (all of) the properties in that set. Our characterization highlights the way in which the clusterings that are output by linkage-based algorithms are different from the clusterings output by other clustering algorithms.

The following is our main result.

THEOREM 2 ([2]). *An clustering function is linkage based if and only if it is local, outer-consistent, outer-rich, and refinement-confined.*

The properties locality, outer-consistency, outer-richness, and refinement-confinement are included in Appendix A. In [1], we show that a similar result holds in the hierarchical clustering setting.

7. REFERENCES

[1] M. Ackerman and S. Ben-David. Discerning Linkage-Based Algorithms Among Hierarchical Clustering Methods. IJCAI, 2011.

[2] M. Ackerman, S. Ben-David, and D. Loker. Characterization of Linkage-based Clustering. COLT, 2010.
[3] M. Ackerman, S. Ben-David, and D. Loker. Towards Property-Based Classification of Clustering Paradigms. NIPS, 2010.
[4] M. Ackerman and S. Ben-David. Clusterability: A Theoretical Study. AISTATS, 2009.
[5] M. Ackerman and S. Ben-David. Measures of Clustering Quality: A Working Set of Axioms for Clustering. NIPS, 2008.
[6] N. Balcan and P. Gupta. Robust Hierarchical Clustering. COLT, 2010.
Proceedings of the 20th annual conference on Learning theory. pp. 20-34, 2007.
[7] A. Blum, N. Nalcan, and S. Vermpala. A Discriminative Framework for Clustering via Similarity Functions. STOC, 2008.
[8] A. Blum, N. Nalcan, and A. Gupta. Approximate Clustering without the Approximation. SODA, 2009.
[9] R. Bosagh Zadeh and S. Ben-David. "A Uniqueness Theorem for Clustering." The 25th Annual Conference on Uncertainty in Artificial Intelligence UAI, 2009.
[10] D. Bryant. On the uniqueness of the selection criterion in neighbor-joining. Journal of Classification 22:3-15, 2005.
[11] S. Epter, M. Krishnamoorthy, and M. Zaki. "Clusterability detection and initial seed selection in large datasets." Technical Report 99-6, Rensselaer Polytechnic Institute, Computer Science Dept., Rensselaer Polytechnic Institute, Troy, NY 12180, 1999.
[12] N. Jardine, R. Sibson, Mathematical Taxonomy Wiley, 1971.
[13] J. Kleinberg. "An Impossibility Theorem for Clustering." Advances in Neural Information Processing Systems (NIPS) 15, 2002.
[14] U. von Luxburg. A Tutorial on Spectral Clustering. Statistics and Computing 17(4): 395-416, 2007
[15] M. Meila. Comparing clusterings: an axiomatic view. Proceedings of the 22nd international conference on Machine learning: 577-584. ACM, 2005.
[16] G.W. Milligan. A Monte Carlo study of thirty internal criterion measures for cluster analysis. Psychometrika 46(2): 187-199, 1981.
[17] R. Ostrovsky, Y. Rabani, L.J. Schulman, and C. Swamy. "The Effectiveness of Lloyd-Type Methods for the k-Means Problem." FOCS '05. 47th Annual IEEE Symposium. Berkeley, CA, USA, pp. 165-176, 2006.
[18] J. Puzicha, T. Hofmann, J. Buhmann. A Theory of Proximity Based Clustering: Structure Detection by Optimization, Pattern Recognition 33, 2000.
[19] L. Vendramin, R.J.G.B. Campello, and E.R. Hruschka. On the comparison of relative clustering validity criteria. Sparks, 2009.
[20] W.E. Wright. A formalization of cluster analysis. Pattern Recognition 5(3):273-282, 1973.
[21] B. Zhang. Dependence of Clustering Algorithm Performance on Clustered-ness of Data. Technical Report, 20010417. Hewlett-Packard Labs, 2001.

Appendix A: Properties of Clustering Functions

In this section we provide formal definitions of the properties discussed throughout this report. We proposed most of these properties in [2] and [3]. A few of the properties come from previous work, we indicate when this is the case.

Locality: Intuitively, a clustering function is local if its behavior on a union of clusters depends only on distances between elements of that union, and is independent of the rest of the domain set.

A clustering function F is *local* if for any clustering C output by F and every subset of clusters, $C' \subseteq C$, $F(\bigcup C', d, |C'|) = C'$.

In other words, for every domain (X, d) and number of clusters, k, if X' is the union of k' clusters in $F(X, d, k)$ for some $k' \leq k$, then, applying F to (X', d) and asking for a k'-clustering, will yield the same clusters that we started with.

Consistency: Consistency, proposed by Kleinberg [13], aims to formalize the preference for clusters that are dense and well-separated. This property requires that the output of a clustering function should remain unchanged after shrinking within-cluster distances and stretching between-cluster distances.

Given a clustering C of some domain (X, d), we say that a distance function d' over X, is (C, d)-*consistent* if $d'(x, y) \leq d(x, y)$ whenever $x \sim_C y$, and $d'(x, y) \geq d(x, y)$ whenever $x \not\sim_C y$. A clustering function F is *consistent* if for every X, d, k, if d' is $(F(X, d, k), d)$-consistent then $F(X, d, k) = F(X, d', k)$.

Kleinberg's original formalization was for general clustering functions, whose only input is the data set (X, d). Consistency for general clustering function is a direction reformulation of the above definition. A general clustering function F is *consistent* if for every X, d, if d' is $(F(X, d), d)$-consistent then $F(X, d) = F(X, d')$.

The following two properties are natural relaxations of consistency. Outer consistency represents the preference for well separated clusters, by requiring that the output of a clustering function not change if clusters are moved away from each other.

Given a clustering C of some domain (X, d), we say that a distance function d' over X, is (C, d)-*outer consistent* if $d'(x, y) = d(x, y)$ whenever $x \sim_C y$, and $d'(x, y) \geq d(x, y)$ whenever $x \not\sim_C y$. A clustering function F is *outer consistent* if for every X, d, k, if d' is $(F(X, d, k), d)$-outer consistent then $F(X, d, k) = F(X, d', k)$.

Inner consistency is defined analogous, representing the preference for placing points that are close together within the same cluster, by requiring that the output of a k-clustering function not change if elements of the same cluster are moved closer to each other. Clearly, consistency implies both outer-consistency and inner-consistency. Note also that if a function is both inner-consistent and outer-consistent then it is consistent.

Richness: These properties require that we be able to obtain any partition of the domain by modifying the distances between elements. Kleinberg's [13] original richness axiom applies to general clustering functions, that map data sets to partitions. A general clustering function F satisfies *richness* if for any sets $X_1, X_2, \ldots, X_k$, there exists a distance function d over $X' = \bigcup_{i=1}^{k} X_i$ so that $F(X', d) = \{X_1, X_2, \ldots, X_k\}$. The following is a natural variation of richness to clustering functions that require the number of clusters, k, to be a part of the input. A clustering function F satisfies *k-richness* if for any sets $X_1, X_2, \ldots, X_k$, there exists a distance function d over $X' = \bigcup_{i=1}^{k} X_i$ so that $F(X', d, k) = \{X_1, X_2, \ldots, X_k\}$.

Outer/Inner Richness: Given k sets, a clustering function satisfies outer richness if there exists some way of setting the between-set distances, without modifying distances within the sets, we can get F to output each of these data sets as a cluster. This corresponds to the intuition that any groups of points, regardless of within distances, can be made into separate clusters.

A clustering function F is *outer-rich* if for every set of domains, $\{(X_1, d_1), \ldots (X_n, d_k)\}$, there exists a distance function $\hat{d}$ over $\bigcup_{i=1}^{n} X_i$ that extends each of the d_i's (for $i \leq k$), such that $F(\bigcup_{i=1}^{k} X_i, \hat{d}, k) = \{X_1, X_2, \ldots, X_k\}$.

Complementary to outer richness, inner richness requires that there be a way of setting distances within sets, without modifying distances between the sets, so that F outputs each set as a cluster. This corresponds to the intuition that between-cluster distances cannot eliminate any partition.

Threshold Richness: Fundamentally, the goal of clustering is to group points that are close to each other, and to separate points that are far apart. Axioms of clustering need to represent these objectives and no set of axioms of clustering can be complete without integrating such requirements. Consistency is the only previous property that aims to formalize these requirements. However, consistency has some counterintuitive implications (as discussed in [5]), and is not satisfied by many common clustering functions. Threshold richness is a property that represents the goal of clustering while being satisfied by common clustering functions.

A clustering function F is *threshold-rich* if for every clustering C of X, there exist real numbers $a < b$ so that for every distance function d over X where $d(x, y) \leq a$ for all $x \sim_C y$, and $d(x, y) \geq b$ for all $x \not\sim_C y$, we have that $F(X, d, |C|) = C$.

Refinement Confinement: A clustering C of X is a *refinement* of clustering C' of X if every cluster in C is a subset of some cluster in C', or, equivalently, if every cluster of C' is a union of clusters of C. A clustering function is *refinement confined* if for every $1 \leq k \leq k' \leq |X|$, $F(X, d, k')$ is a refinement of $F(X, d, k)$.

Order Invariance: Order invariance, proposed by Jardine and Sibson[12], describes clustering functions that are based on the ordering of pairwise distances. A distance function d' of X is an *order invariant modification* of d over X if for all $x_1, x_2, x_3, x_4 \in X$, $d(x_1, x_2) < d(x_3, x_4)$ if and only if $d'(x_1, x_2) < d'(x_3, x_4)$. A clustering function F is *order invariant* if whenever a distance function d' over X is an order invariant modification of d, $F(X, d, k) = F(X, d', k)$ for all k.

Isomorphism Invariance: The following invariance property seems to be an essential part of our understanding of what clustering is. It requires that the output of a clustering function is independent of the labels of the data points. A k-clustering function F is *isomorphism invariant* if whenever $(X, d) \sim (X', d')$, then, for every k, $F(X, d, k)$ and $F(X', d', k)$ are isomorphic clusterings.

Scale Invariance: Scale invariance requires that the output of a clustering function be invariant to uniform scaling of the data. A clustering function F is *scale invariant* if for any data sets (X, d) and (X, d'), if there exists a real number $c > 0$ so that for all $x, y \in X$, $d(x, y) = c\dot{d}'(x, y)$ then for every $1 \leq k \leq |X|$, $F(X, d, k) = F(X, d', k)$. The formulation of scale-invariance in Kleinberg's framework, for general clustering functions do not take the number of clusters as part of the input, is analogues.

Appendix B: Clustering Functions

In this appendix we define some common clustering functions. One such class are linkage-based clustering algorithms, discussed in Section 6. Many clustering algorithms aim to find clusterings with low loss with respect to a specific objective function. An example of such an objective function is Min-Sum, the sum of within-cluster distances, $\sum_{x \sim_C y} d(x, y)$. Every objective function $\mathcal{O}$ has a corresponding clustering function which outputs a clustering that optimizes $\mathcal{O}$. We present some of the most common objective functions below.

Centroid: Following Kleinberg's [13] definition, (k, g)-centroid clustering functions find a set of k "centroids" $\{c_1, \ldots, c_k\} \subseteq X$ so that $\sum_{x \in X} \min_i g(d(x, c_i))$ is minimized. Then k-median is obtained by setting g to the identity function.

k-means: The k-means objective is to find a set of k elements $\{c_1, c_2, \ldots, c_k\}$ in the *underlying space*, so that $\sum_{x \in X} \min_i d(x, c_i)$ is minimized. This formalization assumes that the distance between any element of the domain and point in the underlying space is defined. k-means is often applied in Euclidean space, where the k-means objective is equivalent to $\sum_{C_i \in C} \frac{1}{|C_i|} \sum_{x,y \in C_i} d(x, y)^2$.

Spectral: We discuss two clustering functions from spectral clustering: ratio-cut and normalized-cut [14]. Instead of a distance function, spectral clustering uses a similarity function s. The main difference between a distance function and a similarity function is that higher values represent greater similarity when using similarity functions, while the opposite holds for distance functions. It is easy to reformulate all the properties we discuss to use similarity functions.

Let $\bar{C}_i$ denote the data set X without the points in cluster C_i. For $C_1, C_2 \subseteq X$, let $\text{cut}(C_1, C_2) = \sum_{a \in C_1, b \in C_2} s(a, b)$. The ratio cut objective function is $\sum_1^k \frac{\text{cut}(C_i, \bar{C}_i)}{|C_i|}$, and the normalized cut objective function is $\sum_1^k \frac{\text{cut}(C_i, \bar{C}_i)}{\sum_{a,b \in C_i} s(a,b)}$.