Reliable Clustering with Applications to Data Integration

Sainyam Galhotra
Supervised by: Barna Saha
UMass Amherst
sainyam@cs.umass.edu

ABSTRACT

Today, a number of critical applications that have a significant societal impact are powered by data driven Artificial Intelligence. Given their ubiquity, it is very critical to perform accurate and fair analytics. Entity resolution, community detection and taxonomy construction are some of the building blocks of these applications and for these methods, clustering is one of the common fundamental underlying concept. Therefore, the use of accurate, robust and fair methods for clustering cannot be overstated. This work focuses on these different facets of clustering.

First, we study the problem of clustering in the presence of supervision, specifically aimed at entity resolution. In this setting, we study the robustness and scalability of the methods that leverage supervision through an oracle i.e an abstraction of crowdsourcing.

Second, community detection applications suffer from evaluation in real world scenarios due to lack of ground truth data. We propose a generative model to capture interactions between records that belong to different clusters and devise techniques for efficient cluster recovery.

Third, manifestation of bias in data could arise due to discriminatory treatment of marginalized groups, sampling methods or even measurement errors in the data. We study the impact of this bias on generated clusters and develop techniques that guarantee fair representation from different groups. We prove the noise tolerance of our algorithms and back the theory by demonstrating the efficacy and efficiency on various real world datasets for these applications.

1. INTRODUCTION

With the advances in machine learning and availability of vast amounts of data, Artificial Intelligence based systems are allowed to make autonomous decisions. Already, software makes decisions in who gets a loan [24], hiring [1], self-driving car actions that may lead to property damage or human injury [22], medical diagnosis and treatment [28], and every stage of the criminal justice system including arraignment and sentencing that determine who goes to jail and who is set free [6]. The importance of these decisions makes fairness and quality of the employed algorithms of prime importance.

A number of these real-world applications employ entity resolution, community detection, taxonomy construction and outlier detection as some of their key constituents. Clustering is one of the fundamental techniques that is commonly used to formally study these components. Clustering has been studied for many decades and is considered a challenging task that has evolved over time. In the modern era of big data, the problems of noise, bias and poor quality data have adversely affected the quality of traditional clustering techniques. Along with quality, it is very important to improve their scalability to run on web-scale datasets. Additionally, clustering is generally an unsupervised task and suffers from lack of ground truth data for effective evaluation. There has been a lot of interest in devising generative models to simulate real-world interaction between records of different clusters and benchmark various techniques. In this work, we focus on these different facets of clustering along with its applications towards data integration. Table 1 presents a summary of our contributions.

1.1 Clustering using supervision

Clustering is an intricate problem especially due to the absence of domain knowledge, and the final set of clusters identified using automated techniques can be highly inaccurate and noisy. There has been a lot of recent interest to leverage humans to answer pairwise queries of the form 'do u and v belong to the same optimal cluster?'. Since humans have much more context and domain knowledge, they can answer such queries quite easily. For this reason, many frameworks have been developed to leverage humans (abstracted as an oracle) to perform entity resolution, one of the traditional applications of oracle based clustering techniques in data integration.

Entity Resolution refers to the task of identifying all records that refer to the same entity. Entity resolution is one of the classical data management problems that has been studied since the seminal work of Fellegi and Sunter in 1969 [12]. The explosion of data sources has aggravated the presence of duplicates in a dataset, elevating the importance of Entity Resolution (abbreviated as ER and often referred to as de-duplication). Web-Scale algorithms for de-duplication and organization of data is the need of the hour. ER has evolved from using rule based systems to using human annotators for expert guidance. In traditional settings, the goal of ER

Proceedings of the VLDB 2020 PhD Workshop, August 31st, 2020. Tokyo, Japan.

Table 1: Summary of contributions.

Robust and Scalable	Oracle-based Clustering: Entity Resolution	[16, 14, 17]
	Semantic concept identification and Feature Enrichment	[18]
Generative Model	Geometric Block Model	[20, 19]
Fair and interpretable	Fair Correlation Clustering	[2]
	Interpretable k-center Clustering	[27]

was to match records obtained from two data sources which has now evolved to identify a cluster of records referring to same entity. The heterogeneity of data sources has raised the amount of noise in these datasets and motivated the study of ER scalability. There has been limited work on holistic approaches to identify entities across multiple sources. We develop techniques that are able to resolve entities in datasets with varied cluster distributions and noise levels. To achieve this goal, we make the following contributions.

- **Robustness.** The queries to the oracle can have low accuracy based on their difficulty. Prior oracle-based clustering techniques [29, 30, 15] assumed that all the answers returned by the oracle are correct and hence constructed a spanning tree over the queried edges to identify all the matching pairs. In the absence of noise, this was sufficient due to transitivity (if u, v refer to same entity and v, w refer to same entity then u,w can be inferred as same entity) but it leads to very poor F-score of generated clusters even in case of low error. We propose a cost-effective approach [16, 14] that can be added as an extra-layer to any oracle-based strategy, helping to preserve the performance guarantees of [31, 29, 15] along with high precision. Instead of constructing spanning tree over the records, our approach strengthens all the cuts by constructing sparse graphs with strong connectivity properties. We achieve this with the help of expander graphs [5] and prove precision guarantees of our technique. The error correction layer can be tuned (or even turned off) trading off budget for accuracy, thereby providing flexibility to adapt to different ER applications. In order to efficiently leverage this toolkit, we propose an adaptive technique that changes the connectivity strength of the queried graph based on noise in results and prior similarity of record pairs. We empirically demonstrate that our technique achieves high F-score over different real world datasets.

- **Scalability.** ER is generally preceded by blocking as a pre-processing step to handle large scale datasets. Blocking constitutes the first step that selects sub-quadratic number of record pairs to compare in the subsequent steps. Blocking groups similar records into *blocks* and then selects pairs from the "cleanest" blocks – i.e., those with fewer non-matching pairs – for further comparisons in the pair matching phase. The literature is rich with methods for building and processing blocks [25], but depending on the data, blocking techniques are either (a) too aggressive that they help scale but adversely affect ER accuracy, or (b) too permissive to potentially harm ER efficiency. Due to these limitations, blocking require tuning for each dataset and is one of the most time-consuming components of the pipeline.

  We propose a new methodology of *progressive* blocking [17] that overcomes the above limitations by self-regulating blocking and adapting to the properties of each dataset, with no configuration effort. Our approach performs blocking and matching in tandem, where pair matching results are fed back to the blocking to refine and improve its quality. We demonstrate that our technique achieves the best trade-of between the quality of final results and ER efficiency for a variety of million scale datasets.

As a future work, we are planning to extend oracle-based techniques to perform hierarchical clustering. Hierarchical clustering techniques are very useful to construct taxonomies, analyze phylogenetic trees and construct product catalogs. In this setting, we assume that all the leaf level records are known and the goal is to organize these records in the form of a type-subtype hierarchy. Pairwise oracle query between two leaf level records is not sufficient to construct the hierarchy. Therefore, we consider a triplet query consisting of three records and the oracle identifies the pair of nodes that are closer to each other than the third node. The oracle output provides a local evidence of the hierarchy and is helpful to uncover the structure. One of the key challenges in this line of work is to efficiently identify a small set of queries that can help recover the hierarchy. For a dataset of n records, the total number of possible triplet queries is $O(n^3)$ and enumerating all such queries is impossible for million scale datasets. We leverage pairwise similarities as a guidance to quickly identify the most beneficial triplet queries. Our algorithm maintains a hierarchy of all the processed records and iteratively processes each node with the help of already identified beneficial queries. We show that our technique is able to construct the hierarchy with $O(n \log n)$ queries under reasonable assumptions of the similarity distribution. This work is under progress and we are currently evaluating the quality of our techniques with respect to other baselines.

In addition to oracle based clustering techniques, we are exploring the use of semantic knowledge present in the form of knowledge graphs to identify clusters of web tables and columns that refer to the same concept [18]. Given the scale of data available over the web, the amount of noise and missing information, identifying these clusters is quite challenging. To achieve this goal, we propose an index structure that uses semantic knowledge graphs to quickly identify the distribution of concepts for a particular column. Currently, our index supports text based attributes but does not work for numerical attributes like population, year, age, etc. Identifying clusters of numerical columns requires additional context from the meta-data and other co-occurring columns. We are developing a unified framework to identify semantically coherent clusters of columns and further use these for applications like dataset discovery, feature enrichment, improving search, etc.

2. ABSENCE OF GROUND TRUTH

In this section, we discuss clustering from the lens of community detection over social networks. There are a plethora

of techniques that are used to identify clusters of records referring to same community. However, all these datasets suffer from the scarcity of ground truth data. In order to circumvent this drawback, generative models have been proposed to model the interaction between records of different communities. These models are helpful to benchmark the quality of known clustering techniques to identify clusters.

Stochastic block model (SBM) is one of the most popular random graph model that generalizes the Erdős-Renyi graphs. According to SBM, edges between every pair of nodes are drawn randomly with probability p if the endpoints belong to the same cluster and q if they belong to different clusters. One aspect that SBM does not capture is the 'transitivity rule' (friends having common friends), which is inherent to formation of communities over social networks. Intuitively, if two nodes x, y are connected by an edge and y, z are connected by an edge then it is more likely than not that x, z are connected by an edge. Inspired by this, we proposed the geometric block model [20, 19] that models community formation according to random geometric graphs. One of the key distinction from SBM is that it considers correlated edge formation, capturing the properties of transitivity rule. We empirically validated the model over collaboration networks and co-purchase networks.

We observed that traditional techniques that were developed for cluster recovery in SBM could not be used for the geometric block model. We proposed a simple motif-based counting algorithm to identify clusters and show that it is optimal upto a constant fraction. We tested the effectiveness of our algorithm to recover clusters over various real-world and synthetic datasets.

3. FAIRNESS AND INTERPRETABILITY

There are a countless number of examples where the use of biased systems have led to disastrous consequences. Clustering techniques are used in various applications like team formation and community detection which have societal impact. Given their importance, there has been little work on improving the fairness and interpretability of these algorithms. We consider different clustering techniques and devise scalable methods to improve their fairness and interpretability.

Correlation Clustering. Correlation clustering, introduced by Bansal, Blum and Chawla in 2004 [7], has received tremendous attention in the past decade. The problem is NP-complete and a series of follow-up work has resulted in better approximation ratio, generalization to weighted graphs, etc. [4, 9, 10]. This problem captures a wide range of applications including clustering gene expression patterns [8, 23], and the aggregation of inconsistent information [13].

Chierichetti et al. [11] extended the notion of disparate impact to k-center and k-median objectives, and studied these problems for the case of two groups. Their result was later generalized to multiple groups by Rösner and Schmidt [26]. We generalize the notion of disparate impact [2] to correlation clustering for multiple colors and our goal is to make sure that the distribution of colors in each cluster is identical to the global distribution. Additionally, we extend the model introduced by Ahmadian et al. [3] on k-center to correlation clustering to ensure that no color is over or under represented in each cluster.

More formally, our fairness-aware variant of correlation clustering [7] identifies clusters while ensuring equal distribution of demographics. Our algorithm proceeds in two steps. In the first step it identifies a matching between nodes of different colors to construct small clusters that satisfy fairness constraints. In the second step it chooses representative nodes (one from each matched clusters) and employs traditional correlation clustering algorithm to identify the final set of clusters. We prove that our algorithm identifies clusters within a constant factor approximation of the optimal solution. We further relax the equal distribution constraint and extend our algorithm for a lower and upper bound constraint on the number of nodes of each color in a cluster. To further instill trust in the data, we explore multi-objective clustering algorithms to generate explainable clusters with minimal loss in the clustering objective.

Interpretable Clustering. Clustering techniques are expected to be inherently interpretable as the goal is to group similar nodes together. However, with the increase in number of features for each record, the generated clusters can have poor interpretability. In our work [27], we measure interpretability in terms of the homogeneity of nodes in a cluster with respect to the features of interest for the end-user. We consider the k-center clustering objective and develop techniques to achieve β-interpretability (for a given parameter β) with respect to features of interest. The choice of β determines the trade-off between clustering objective and interpretability.

Multi-Objective Clustering. With the increased societal impact of clustering techniques, the importance of considering additional constraints like fairness, diversity, interpretability and efficiency has increased. This has motivated the study of multi-objective clustering techniques focused towards these objectives. Existing techniques that support multi-objective clustering either leverage a scalarization function, which combines the multiple objectives into a single objective, or find clusters in parallel for each objective and combine the results using different approaches such as a fitness function. Such techniques lose theoretical guarantees with respect to any of the considered objectives. In [21], we consider a lexicographic multi-objective framework where the optimization objectives are lexicographically ordered and our optimization algorithm follows the same preference. In this setting, the goal is to prioritize primary clustering objectives over ancillary objectives. To further simulate different scenarios, our model uses a slack value to improve the quality on secondary objectives and allows minor deviations of the primary objective from its optimal value. Our algorithm processes the different objectives in the order of their preference and generates final clustering. Incase of any violation of clustering objectives, local search techniques are employed to satisfy the corresponding slack values.

4. CONCLUSION AND FUTURE WORK

In this work, we have studied the different facets of clustering focussing on robustness, scalability, generative modelling, fairness and interpretability of flat clustering algorithms. We demonstrated the effectiveness of our techniques to perform clustering with applications towards entity resolution, community detection and other societal issues of bias and discrimination. We study entity resolution from the perspective of using oracles as an abstraction of humans to answer pairwise queries and discuss the importance of scalable techniques for web-scale datasets. In community

detection, we study generative models to simulate interaction between records of different clusters. Additionally, we study traditional clustering techniques along with fairness and interpretability constraints. As a future work, we are working towards extending our work to consider these different facets for hierarchical clustering for applications like taxonomy construction, knowledge graph construction, data organization and team formation.

5. ACKNOWLEDGEMENT

The authors would like to thank all the contributors, Barna Saha (advisor), Donatella Firmani, Divesh Srivastava, Arya Mazumdar, Soumyabrata Pal, Saba Ahmadi, Roy Schwartz, Sandhya Saisubramanian, Shlomo Zilberstein, Udayan Khurana, Oktie Hassanzadeh and Kavitha Srinivas.

6. REFERENCES

[1] Are ai hiring programs eliminating bias or making it worse? *Forbes*.

[2] S. Ahmadi, S. Galhotra, B. Saha, and R. Schwartz. Fair correlation clustering, 2020.

[3] S. Ahmadian, A. Epasto, R. Kumar, and M. Mahdian. Clustering without over-representation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 267–275, 2019.

[4] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)*, 55(5):1–27, 2008.

[5] N. Alon and J. H. Spencer. *The probabilistic method*. John Wiley & Sons, 2004.

[6] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. *ProPublica*, May 23, 2016.

[7] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine learning*, 56(1-3), 2004.

[8] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of computational biology*, 6(3-4):281–297, 1999.

[9] M. Charikar, V. Guruswami, and A. Wirth. Clustering with qualitative information. *Journal of Computer and System Sciences*, 71(3):360–383, 2005.

[10] S. Chawla, K. Makarychev, T. Schramm, and G. Yaroslavtsev. Near optimal lp rounding algorithm for correlationclustering on complete and complete k-partite graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 219–228, 2015.

[11] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5029–5037. Curran Associates, Inc., 2017.

[12] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[13] V. Filkov and S. Skiena. Integrating microarray data by consensus clustering. *International Journal on Artificial Intelligence Tools*, 13(04):863–880, 2004.

[14] D. Firmani, S. Galhotra, B. Saha, and D. Srivastava. Robust entity resolution using a crowdoracle. 2018.

[15] D. Firmani, B. Saha, and D. Srivastava. Online entity resolution using an oracle. *PVLDB*, 9(5):384–395, 2016.

[16] S. Galhotra, D. Firmani, B. Saha, and D. Srivastava. Robust entity resolution using random graphs. In *SIGMOD*, 2018.

[17] S. Galhotra, D. Firmani, B. Saha, and D. Srivastava. Efficient and effective er with progressive blocking, 2020.

[18] S. Galhotra, U. Khurana, O. Hassanzadeh, K. Srinivas, H. Samulowitz, and M. Qi. Automated feature enhancement for predictive modeling using external knowledge. *ICDM*, 2019.

[19] S. Galhotra, A. Mazumdar, S. Pal, and B. Saha. Connectivity in random annulus graphs and the geometric block model. *CoRR*, abs/1804.05013, 2018.

[20] S. Galhotra, A. Mazumdar, S. Pal, and B. Saha. The geometric block model. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[21] S. Galhotra, S. Saisubramanian, and S. Zilberstein. Lexicographically ordered multi-objective clustering. *arXiv preprint arXiv:1903.00750*, 2019.

[22] N. J. Goodall. Can you program ethics into a self-driving car? *IEEE Spectrum*, 53(6):28–58, June 2016.

[23] J. Guo, F. Hüffner, C. Komusiewicz, and Y. Zhang. Improved algorithms for bicluster editing. In M. Agrawal, D. Du, Z. Duan, and A. Li, editors, *Theory and Applications of Models of Computation*, pages 445–456, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[24] P. Olson. The algorithm that beats your bank manager. *CNN Money*, March 15, 2011.

[25] G. Papadakis, J. Svirsky, A. Gal, and T. Palpanas. Comparative analysis of approximate blocking techniques for entity resolution. *Proceedings of the VLDB Endowment*, 9(9):684–695, 2016.

[26] C. Rösner and M. Schmidt. Privacy preserving clustering with constraints. *arXiv preprint arXiv:1802.02497*, 2018.

[27] S. Saisubramanian, S. Galhotra, and S. Zilberstein. Balancing the tradeoff between clustering value and interpretability. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 351–357, New York, NY, USA, 2020.

[28] E. Strickland. Doc bot preps for the O.R. *IEEE Spectrum*, 53(6):32–60, June 2016.

[29] N. Vesdapunt, K. Bellare, and N. Dalvi. Crowdsourcing algorithms for entity resolution. *PVLDB*, 7(12):1071–1082, 2014.

[30] J. Wang, T. Kraska, M. J. Franklin, and J. Feng. Crowder: Crowdsourcing entity resolution. *PVLDB*, 5(11):1483–1494, 2012.

[31] J. Wang, G. Li, T. Kraska, M. J. Franklin, and J. Feng. Leveraging transitive relations for crowdsourced joins. In *SIGMOD Conference*, 2013.