Clustering with a faulty oracle

Kasper Green Larsen
larsen@cs.au.dk
Aarhus University

Michael Mitzenmacher
michaelm@eecs.harvard.edu
Harvard University

Charalampos E. Tsourakakis
ctsourak@bu.edu
Boston University

ABSTRACT

Clustering, i.e., finding groups in the data, is a problem that permeates multiple fields of science and engineering. Recently, the problem of clustering with a noisy oracle has drawn attention due to various applications including crowdsourced entity resolution [33], and predicting signs of interactions in large-scale online social networks [20, 21]. Here, we consider the following fundamental model for two clusters as proposed by Mitzenmacher and Tsourakakis [28], and Mazumdar and Saha [25]; there exist n items, belonging to two unknown groups. We are allowed to query any pair of nodes whether they belong to the same cluster or not, but the answer to the query is corrupted with some probability $0 < q < \frac{1}{2}$. Let $1 > \delta = 1 - 2q > 0$ be the bias.

In this work, we provide a polynomial time algorithm that recovers all signs correctly with high probability in the presence of noise with $O(\frac{n \log n}{\delta^2} + \frac{\log^2 n}{\delta^6})$ queries. This is the best known result for this problem for all but tiny δ, improving on the current state-of-the-art due to Mazumdar and Saha [25].

CCS CONCEPTS

• **Mathematics of computing → Graph algorithms**; • **Theory of computation → Graph algorithms analysis**.

KEYWORDS

clustering, active learning, randomized algorithms

ACM Reference Format:
Kasper Green Larsen, Michael Mitzenmacher, and Charalampos E. Tsourakakis. 2020. Clustering with a faulty oracle. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3366423.3380045

1 INTRODUCTION

Clustering is a central problem in data science with a rich history; hundreds of algorithms have been published on the topic. Certain popular algorithms that have inaugurated lines of research include k-means and k-means++ (e.g., [3, 4, 19]), mixture models (e.g., [26]), spectral clustering (e.g., [2, 31]), correlation clustering (e.g., [5]), graph clustering methods (e.g., [7]). Despite the long research history, clustering remains an active area of research. Part of the reason why this is true is that recent advances in technologies, data availability etc. motivate new variants of clustering problems. In this work we focus on clustering with a faulty oracle. This particular clustering variant is motivated by various applications including the humans-in-the-loop approach to the entity resolution problem, and predicting signed edges in large-scale online social networks, see [25]. Furthermore, this clustering variant has various interesting connections with other lines of research on clustering including the stochastic model, and correlation clustering that we discuss in section 2. We define the model that we study in the following.

Model. Let $V = [n]$ be the set of n items that belong to two clusters. Set $\sigma : V \to \{-1, +1\}$, and let $R = \{v \in V(G) : \sigma(v) = -1\}$ and $B = \{v \in V(G) : \sigma(v) = +1\}$ be the sets/groups of red and blue nodes respectively, where $0 \leq |R| \leq n$. For any pair of nodes $\{u, v\}$ define $\tau(u, v) = \sigma(u)\sigma(v) \in \{\pm 1\}$ (i.e., $\tau(u, v) = -1$, if u is reported to be in the different cluster than v). The coloring function σ is unknown and we wish to recover the two sets R, B by querying pairs of items. (We need not recover the labels, just the clusters.) Let $\eta_{u,v} \in \{\pm 1\}$ be iid noise in the edge observations, with $\mathbb{E}\left[\eta_{u,v}\right] = \delta$ for all pairs $u, v \in V$. The oracle returns

$$\tilde{\tau}(u, v) = \sigma(u)\sigma(v)\eta_{u,v}.$$

Equivalently, for each query we receive the correct answer with probability $1 - q = \frac{1}{2} + \frac{\delta}{2}$, where $q > 0$ is the corruption probability. Our goal is answer the following question.

Problem 1.1. Can we recover the clusters *efficiently* with high probability by performing a *small number of queries*?

The constraint of querying a pair of nodes *only once* in the presence of noise appears not only in settings where a repeated query is constrained to give the same answer but naturally in more complex settings. For example, the entity resolution problem is a classic problem in data management that aims to identify and group records that refer to the same entity. Recently, crowdsourcing platforms like Amazon Mechanical Turk are used to attack this problem by presenting workers with questions of the form "do these two items represent the same entity?". The goal is to solve the entity resolution problem while minimizing the monetary cost of the process. The workers' answers are not always reliable. This can be modeled using the noisy oracle model that we study. Once deciding on whether a given pair of items refers to the same entity or not by looking at the workers' answers, no more queries are typically performed. Interestingly, it has also been observed empirically that repeated querying does not help much in reducing errors [24, 25, 33].

Main results. Our main theoretical result shows that we can recover the two clusters (R, B) with high probability[1] in polynomial time. Specifically, our proposed algorithm runs in time $O(\frac{n \log n}{\delta^2} + \frac{\log^3 n}{\delta^8})$. Our result is stated as the next theorem.

Theorem 1.2. *There exists a polynomial algorithm with query complexity $O(\frac{n \log n}{\delta^2} + \frac{\log^2 n}{\delta^6})$ that returns both clusters of V whp.*

[1] An event A_n holds with high probability (*whp*) if $\lim_{n \to +\infty} \mathbf{Pr}[A_n] = 1$.

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380045

Our algorithm improves the current state-of-the-art due to Mazumdar and Saha [25]. Specifically, their information theoretical optimal algorithm that performs $O(\frac{n\log n}{\delta^2})$ queries requires quasi-polynomial runtime and is unlikely to be improved assuming the planted clique conjecture. On the other hand, their efficient polytime algorithms require $O(\frac{n\log n}{\delta^4})$ queries. Our algorithm is optimal for all but tiny δ, i.e., as long as the first term $\frac{n\log n}{\delta^2}$ dominates (asymptotically) the second term $\frac{\log^2 n}{\delta^6}$.

Roadmap. In Section 2 we briefly overview related work, and in Section 3 we present and analyze Pythia2Truth, our proposed algorithm. Its name is inspired by Greek mythology; Pythia was an oracle known to give ambiguous answers to queries. We conclude our short paper with an interesting open problem in Section 4.

2 RELATED WORK

Clustering with Noisy Queries. Closest to our work lies the recent work of Mazumdar and Saha [25]. Specifically, the authors study Problem 1.1 in [25] as well, as well as the more general version where the number of clusters is $k \geq 3$. Each oracle query provides a noisy answer on whether two nodes belong to the same cluster or not. They provide an algorithm that performs $O(\frac{nk\log n}{\delta^2})$ queries, recovers all clusters of size $\Omega(\frac{\log n}{\delta^2})$ where k is the number of clusters, but whose runtime is quasi-polynomial hence impractical, and unlikely to be improved under the planted clique hardness assumption. They also design a computationally efficient algorithm that runs in $O(n\log n + k^6)$ time and performs $O(\frac{nk^2\log n}{\delta^4})$ queries. Finally, for $k = 2$ they provide a non-adaptive algorithm that performs $O(\frac{n\log n}{\delta^4})$ and runs in $O(n\log n)$ time. Previously, sub-optimal results had been obtained by Mitzenmacher and Tsourakakis [28]. It is worth outlining that recovering combinatorial structures using noisy queries is an important problem in theoretical computer science [8, 11, 12, 16, 17, 29].

Correlation Clustering. Bansal et al. [5] studied Correlation Clustering: given an undirected signed graph partition the nodes into clusters so that the total number of disagreements is minimized. This problem is NP-hard [5, 32]. Here, a disagreement can be either a positive edge between vertices in two clusters or a negative edge between two vertices in the same cluster. Note that in Correlation Clustering the number of clusters is not specified as part of the input. 2-Correlation-Clustering refers to the case when the number of clusters is constrained to be at most two.

We remark that the notion of *imbalance* studied by Harary is the 2-Correlation-Clustering cost of the signed graph. Mathieu and Schudy initiated the study of noisy correlation clustering [23]. They develop various algorithms when the graph is complete, both for the cases of a random and a semi-random model. Later, Makarychev, Makarychev, and Vijayaraghavan proposed an algorithm for graphs with $O(n\text{poly}\log n)$ edges under a semi-random model [22]. For more information on Correlation Clustering see the recent survey by Bonchi et al. [6].

Planted bisection model. The following well-studied bisection model is closely connected to our model. Suppose that there are two groups (clusters) of nodes. A graph is generated as follows: the edge probabilities are p within each cluster, and $q < p$ across the clusters. The goal is to recover the two clusters given such a graph. If the two clusters are balanced, i.e., each cluster has $O(n)$ nodes, then one can recover the clusters *whp*, see [1, 27, 34]. Hajek, Wu, and Xu proved that when each cluster has $n/2$ nodes (perfect balance), the average degree has to scale as $\frac{\log n}{(\sqrt{1-q}-\sqrt{q})^2}$ for exact recovery [18]. Also, they showed that using semidefinite programming (SDP) exact recovery is achievable at this threshold [18]. Notice that if (i) we have two balanced clusters, and (ii) we remove all negative edges from a signed graph generated according to our model, then one can apply such techniques to recover the clusters. We observe that when $\delta \to 0$ the lower bound of Hajek et al. scales as $O(\frac{\log n}{\delta^2})$.

Other Techniques. Chen et al. [13, 14] consider our model, and provide a method that can reconstruct the clustering for random binomial graphs with $O(n\text{poly}\log n)$ edges. Their method exploits low rank properties of the cluster matrix, and requires certain conditions, including conditions on the imbalance between clusters, see [14, Theorem 1, Table 1]. Their method is based on a convex relaxation of a low rank problem. Mazumdar and Saha similarly study clustering with an oracle in the presence of side information, such as a Jaccard similarity matrix [24]. Cesa-Bianchi et al. [10] take a learning-theoretic perspective on the problem of predicting signs. They use the correlation clustering objective as their learning bias, and show that the risk of the empirical risk minimizer is controlled by the correlation clustering objective. Chiang et al. point out that the work of Candès and Tao [9] can be used to predict signs of edges, and also provide various other methods, including singular value decomposition based methods, for the sign prediction problem [15]. The incoherence is the key parameter that determines the number of queries, and is equal to the group imbalance $\tau = \max_{\text{cluster } C} \frac{n}{|C|}$. The number of queries needed for exact recovery under our model is $O(\tau^4 n\log^2 n)$, which is prohibitive when clusters are even slightly imbalanced.

3 PROPOSED METHOD

We describe our proposed algorithm Pythia2Truth that achieves the guarantees of Theorem 1.2. The algorithm arbitrarily chooses two sets $A, B \subseteq V$ such that $|A| = O(\frac{\log n}{\delta^2})$ and $|B| = O(\frac{\log n}{\delta^4})$. Then, it performs all possible queries between A, B. The total number of queries at this step is $O(\frac{\log^2 n}{\delta^6})$. The algorithm then uses the set of labels $\{\tilde{\tau}(a,b), \tilde{\tau}(a',b)\}_{b\in B}$ to make a guess $\bar{\tau}(a,a')$ for $\tau(a,a')$ for each pair $a, a' \in A$. This works as follows: for any given pair $\{a, a'\}$ each b casts a vote $vote(a,a',b)$. Specifically, $vote(a,a',b) = +1$ if $\tilde{\tau}(a,b) = \tilde{\tau}(a',b)$, and $vote(a,a',b) = -1$ if $\tilde{\tau}(a,b) \neq \tilde{\tau}(a',b)$. The prediction $\bar{\tau}(a,a')$ is $+1$ if the majority of votes $\{vote(a,a',b)\}_{b\in B}$ is $+1$, and -1 otherwise.

The aforementioned steps ensure that $\bar{\tau}(a,a') = \tau(a,a')$ for all pairs $a, a' \in A$ *whp*. Clearly, there exist at least $\Omega(\frac{\log n}{\delta^2})$ nodes from at least one of the two clusters. This set of nodes is found by finding the largest connected component (that is actually a clique) of the graph induced by the positive edges in A. This set C serves as a *seed set*. For each node $u \notin C$ we perform all queries (u, c) for each $c \in C$. If the majority of the oracle answers is $+1$ then we add u in C. The procedure outputs C and its complement as the true clusters.

Algorithm 1 Pythia2Truth(V)

Choose arbitrarily $A, B \subseteq V$ two disjoint sets of nodes, such that $|A| = \frac{48 \log n}{\delta^2}$, and $|B| = \frac{24 \log n}{\delta^4}$.
Perform all $\Theta(\frac{\log^2 n}{\delta^6})$ queries among A, B.
for each pair $a, a' \in A$ **do**
  $\text{counter}_{a,a'} \leftarrow 0$
  for each $b \in B$ **do**
    if $\tilde{\tau}(a,b) = \tilde{\tau}(a',b)$ **then**
      $\text{counter}_{a,a'} \leftarrow \text{counter}_{a,a'} + 1$
    end if
  end for
  if $\text{counter}_{a,a'} \geq \frac{|B|}{2}$ **then**
    $\bar{\tau}(a,a') = +1$
  else
    $\bar{\tau}(a,a') = -1$
  end if
end for
Remove the negative edges from A, and let C be the largest clique
for each $u \in V \backslash C$ **do**
  Perform all queries (u, c) for $c \in C$
  if the majority of answers is + **then**
    $C \leftarrow C \cup \{u\}$
  end if
end for
return $(C, V \backslash C)$

Now we prove the correctness of our proposed algorithm. First, we prove the following lemma.

Lemma 3.1. *Let $S \subseteq V$ such that $|S| = \frac{24 \log n}{\delta^4}$. Consider any pair of nodes $u, v \in V \backslash S$, and let $\bar{\tau}(u,v) = majority(\{\tilde{\tau}(u,s) \cdot \tilde{\tau}(v,s)\}_{s \in S})$. Then, $\bar{\tau}(u,v) = \tau(u,v)$ with probability at least $1 - \frac{1}{n^3}$.*

Proof. Consider any pair of nodes $u, v \in V \backslash S$, and let $X_s(u,v)$ be an indicator random variable for $s \in S$ that is equal to 1 if the product $\tilde{\tau}(u,s) \cdot \tilde{\tau}(v,s)$ of the two noisy labels $\tilde{\tau}(u,s), \tilde{\tau}(v,s)$ is the true label $\tau(u,v)$. Then, $\mathbf{Pr}[X_s = 1] = (1-q)^2 + q^2 = \frac{1+\delta^2}{2}$. For notation simplicity let $p = \mathbf{Pr}[X_s = 1]$. Also, we define $X(u,v) = \sum_{s \in S} X_s(u,v)$. Notice that $\bar{\tau}(u,v) = \tau(u,v)$ iff $X(u,v) \geq \frac{|S|}{2}$. Using Chernoff bounds [30], we obtain that the probability of misclassification is bounded by

$$\begin{aligned}
\mathbf{Pr}\left[X(u,v) < \frac{|S|}{2}\right] &= \mathbf{Pr}\left[X(u,v) < \frac{p|S|}{2p}\right] \\
&= \mathbf{Pr}\left[X(u,v) < \left(1 - (1 - \frac{1}{2p})\right)p|S|\right] \\
&\leq \exp\Big(-\frac{(2p-1)^2}{8p^2}\frac{24\log n}{\delta^4}p\Big) \\
&= \exp\Big(-\frac{\delta^4}{4(1+\delta^2)}\frac{24\log n}{\delta^4}\Big) < \frac{1}{n^3}.
\end{aligned}$$

□

A straight-forward corollary of lemma 3.1 derived by taking a union bound over all pairs of nodes in $V \backslash S$ is that our algorithm predicts the labels of all such interactions correctly *whp*. Using lemma 3.1 we are also able to prove the correctness of our Algorithm.

Proof of Theorem 1.2. Using lemma 3.1 by setting $S = B$ we obtain that all pairwise interactions within the set A are correctly labeled with high probability. By the pigeonhole principle, since $|A| = \frac{48 \log n}{\delta^2}$, one of the two clusters has at least $\frac{24 \log n}{\delta^2}$ nodes in A. This set can easily be found: since within A all labels $\bar{\tau}(a,a')$ are equal to $\tau(a,a')$, for $a, a' \in A$, disregarding the negative labels $\bar{\tau}(a,a')$ will result in at most two connected cliques. We can find the largest such clique in $O(|A|)$ time (since one step of BFS finds all other nodes). Let C be the corresponding set of nodes.

Let $u \in V \backslash C$. We perform all possible $|C|$ queries between u and C, and we decide that u belongs to C if the majority of the oracle answers is +1. Define $X_c(u)$ to be an indicator random variable that is equal to 1 if the oracle answer for the pair $\{u, c\}$ is correct, and 0 otherwise. Let $X(u) = \sum_{c \in C} X_c(u)$ be the random variable distributed according to $Bin(|C|, 1-q)$. The probability of failure is bounded by

$$\begin{aligned}
\mathbf{Pr}\left[X(u) < \frac{|C|}{2}\right] &= \mathbf{Pr}\left[X(u) < \left(1 - (1 - \frac{1}{2(1-q)})\right)(1-q)|C|\right] \\
&\leq \exp\Big(-\frac{\delta^2}{2(1+\delta)^2}\frac{24\log n}{\delta^2}\frac{1+\delta}{2}\Big) < \frac{1}{n^3}.
\end{aligned}$$

By combining the above results, and a union bound our proposed algorithm succeeds *whp* to recover both clusters. □

The total running time of our method is $O\Big(\underbrace{\binom{\frac{48\log n}{\delta^2}}{2}\frac{24\log n}{\delta^4}}_{\text{classify all pairs in } A} + \underbrace{\frac{48\log n}{\delta^2}}_{\text{find largest clique}} + \underbrace{\frac{n\log n}{\delta^2}}_{\text{decide for the rest}}\Big)$ that simplifies to the total running time of $O(\frac{n\log n}{\delta^2} + \frac{\log^3 n}{\delta^8})$.

4 CONCLUSION

An interesting open problem is to achieve optimal query complexity $O(\frac{n\log n}{\delta^2})$ in time linear in the number of queries. In other words, can we remove the $\frac{\log^2(n)}{\delta^6}$ term from our query complexity?

Another open problem relates to the extension of our result to k clusters. Specifically, our clustering model naturally extends to the case where there are more than two clusters [25]. In this case the set V of n items belong to k clusters. When we query the pair of nodes $\{u, v\}$ we obtain a noisy answer on whether u, v belong to the same cluster or not. Can we design a query-optimal, time-efficient algorithm that performs $O(\frac{kn\log n}{\delta^2})$ queries for all $0 < \delta < 1$?

REFERENCES

[1] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. 2016. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory* 62, 1 (2016), 471–487.

[2] Noga Alon and Vitali D Milman. 1985. $\lambda 1$, isoperimetric inequalities for graphs, and superconcentrators. *Journal of Combinatorial Theory, Series B* 38, 1 (1985), 73–88.

[3] David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 1027–1035.

[4] Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii. 2012. Scalable k-means++. *Proceedings of the VLDB Endowment* 5, 7 (2012), 622–633.

[5] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation clustering. *Machine Learning* 56, 1-3 (2004), 89–113.

[6] Francesco Bonchi, David Garcia-Soriano, and Edo Liberty. 2014. Correlation clustering: from theory to practice.. In *KDD*. 1972.

[7] Ulrik Brandes, Marco Gaertler, and Dorothea Wagner. 2003. Experiments on graph clustering algorithms. In *European Symposium on Algorithms*. Springer, 568–579.

[8] Mark Braverman and Elchanan Mossel. 2008. Noisy sorting without resampling. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 268–276.

[9] Emmanuel J Candès, Justin Romberg, and Terence Tao. 2006. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory* 52, 2 (2006), 489–509.

[10] Nicolo Cesa-Bianchi, Claudio Gentile, Fabio Vitale, Giovanni Zappella, et al. 2012. A Correlation Clustering Approach to Link Classification in Signed Networks.. In *COLT*. 34–1.

[11] Moses Charikar, Ronald Fagin, Venkatesan Guruswami, Jon Kleinberg, Prabhakar Raghavan, and Amit Sahai. 2002. Query strategies for priced information. *J. Comput. System Sci.* 64, 4 (2002), 785–819.

[12] Yuxin Chen and Emmanuel J Candès. 2018. The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Communications on Pure and Applied Mathematics* 71, 8 (2018), 1648–1714.

[13] Yudong Chen, Ali Jalali, Sujay Sanghavi, and Huan Xu. 2014. Clustering partially observed graphs via convex optimization. *Journal of Machine Learning Research* 15, 1 (2014), 2213–2238.

[14] Yudong Chen, Sujay Sanghavi, and Huan Xu. 2012. Clustering sparse graphs. In *Advances in neural information processing systems*. 2204–2212.

[15] Kai-Yang Chiang, Cho-Jui Hsieh, Nagarajan Natarajan, Inderjit S Dhillon, and Ambuj Tewari. 2014. Prediction and clustering in signed networks: a local to global perspective. *Journal of Machine Learning Research* 15, 1 (2014), 1177–1213.

[16] Kasper Green Larsen, Michael Mitzenmacher, and Charalampos E Tsourakakis. 2019. Optimal Learning of Joint Alignments with a Faulty Oracle. *arXiv preprint arXiv:1909.09912* (2019).

[17] Anupam Gupta and Amit Kumar. 2001. Sorting and selection with structured costs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*. IEEE, 416–425.

[18] Bruce Hajek, Yihong Wu, and Jiaming Xu. 2016. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory* 62, 5 (2016), 2788–2797.

[19] Anil K Jain. 2010. Data clustering: 50 years beyond K-means. *Pattern recognition letters* 31, 8 (2010), 651–666.

[20] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World Wide Web (WWW)*. ACM, 641–650.

[21] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Signed networks in social media. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 1361–1370.

[22] Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. 2015. Correlation clustering with noisy partial information. In *Proceedings of the Conference on Learning Theory (COLT)*, Vol. 6. 12.

[23] Claire Mathieu and Warren Schudy. 2010. Correlation clustering with noisy input. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 712–728.

[24] Arya Mazumdar and Barna Saha. 2016. Clustering Via Crowdsourcing. *arXiv preprint arXiv:1604.01839* (2016).

[25] Arya Mazumdar and Barna Saha. 2017. Clustering with noisy queries. In *Advances in Neural Information Processing Systems*. 5790–5801.

[26] Geoffrey J McLachlan and Kaye E Basford. 1988. *Mixture models: Inference and applications to clustering*. Vol. 84. M. Dekker New York.

[27] Frank McSherry. 2001. Spectral partitioning of random graphs. In *Proceedings. 42nd IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE, 529–537.

[28] Michael Mitzenmacher and Charalampos E Tsourakakis. 2016. Predicting Signed Edges with $O(n^{1+o(1)})$ Queries. *arXiv preprint arXiv:1609.00750* (2016).

[29] Michael Mitzenmacher and Charalampos E Tsourakakis. 2018. Joint alignment from pairwise differences with a noisy oracle. In *International Workshop on Algorithms and Models for the Web-Graph*. Springer, 59–69.

[30] Michael Mitzenmacher and Eli Upfal. 2005. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge university press.

[31] Andrew Y Ng, Michael I Jordan, and Yair Weiss. 2002. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*. 849–856.

[32] Ron Shamir, Roded Sharan, and Dekel Tsur. 2004. Cluster graph modification problems. *Discrete Applied Mathematics* 144, 1 (2004), 173–182.

[33] Vasilis Verroios and Hector Garcia-Molina. 2015. Entity resolution with crowd errors. In *IEEE 31st International Conference on Data Engineering (ICDE)*. IEEE, 219–230.

[34] Van Vu. 2014. A simple SVD algorithm for finding hidden partitions. *arXiv preprint arXiv:1404.3918* (2014).