# Resource-bounded Information Gathering for Correlation Clustering

Pallika Kanani, Andrew McCallum

University of Massachusetts, Amherst

**Abstract.** We present a new class of problems, called *resource-bounded information gathering for correlation clustering*. Our goal is to perform correlation clustering under circumstances in which accuracy may be improved by augmenting the given graph with additional information. This information is obtained by querying an external source under resource constraints. The problem is to develop the most effective query selection strategy to minimize some loss function on the resulting partitioning. We motivate the problem using an entity resolution task.

## 1 Problem Definition

The standard correlation clustering problem on a graph with real-valued edge weights is as follows: there exists a fully connected graph $G(V, E)$ with $n$ nodes and edge weights, $w_{ij} \in [-1, +1]$. The goal is to partition the vertices in $V$ by minimizing the inconsistencies with the edge weights [1]. That is, we want to find a partitioning that maximizes the objective function $\mathcal{F} = \sum_{ij} w_{ij} f(i, j)$, where $f(i, j) = 1$ when $v_i$ and $v_j$ are in the same partition and $-1$ otherwise.

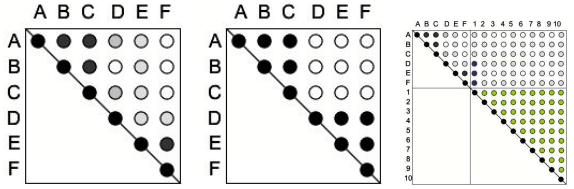
Now consider a case in which there exists some "true" partitioning $\mathcal{P}$, and the edge weights $w_{ij} \in [-\infty, +\infty]$ are drawn from a random distribution (noise model) that is correlated with whether or not edge $e_{ij} \in E$ is cut by a partition boundary. The goal is to find an approximate partitioning, $\mathcal{P}_a$, of $V$ into an unknown number of $k$ partitions, such that $\mathcal{P}_a$ is as 'close' to $\mathcal{P}$ as possible. There are many different possible measures of closeness to choose from. Let $\mathcal{L}(\mathcal{P}, \mathcal{P}_a)$ be some arbitrary loss function. If no additional information is available, then we could simply find a partitioning that optimizes $\mathcal{F}$ on the given weights.

In this paper, we consider settings in which we may issue queries for additional information to help us reduce loss $\mathcal{L}$. Let $G_0(V_0, E_0)$ be the original graph. Let $\mathcal{F}_0$ be the objective function defined over $G_0$. Our goal is to perform correlation clustering and optimize $\mathcal{F}_0$ with respect to the true partitioning of $G_0$. We can augment the graph with additional information using two alternative methods: (1) updating the weight on an existing edge, (2) adding a new vertex and edges connecting it to existing vertices. We can obtain this additional information by querying a (possibly adversarial) oracle using two different types of queries. In the first method, we use query of type Q1, which takes as input edge $e_{ij}$ and returns a new edge weight $w'_{ij}$, where $w'_{ij}$ is drawn from a different distribution that has higher correlation with the true partitioning $\mathcal{P}$.

In the second method, we can expand the graph $G_0$, by adding a new set of vertices, $V_1$ and the corresponding new set of edges, $E_1$ to create a larger, fully connected graph, $G'$. Although we are not interested in partitioning $V_1$, we hypothesize that partitioning $G'$ would improve the optimization of $\mathcal{F}_\prime$ on $G_0$ due to transitivity of partition membership. In this case, given resource constraints, we must select $V'_s \subset V_1$ to add to the graph. These can be obtained by second type of query, Q2, which takes as input $(V_0, E_0)$ and returns a subset $V'_s \subset V_1$. Note that the additional nodes obtained as a result of the queries of type Q2 help by inducing a new, and presumably more accurate partitioning on the nodes of $G_0$. Fig. 1 illustrates the result of these queries.

However, there exist many possible queries of type Q1 and Q2, each with an associated cost. There is also a cost for performing computation on the additional information. Hence, we need an efficient way to select and order queries under the given resource constraints.

Formally, we define the problem of *resource-bounded information gathering for correlation clustering* as follows. Let $c(q)$ be the cost associated with a query $q \in Q1 \cup Q2$. Let $b$ be the total budget on queries and computation. Find distinct queries $q_1, q_2, .....q_m \in Q1 \cup Q2$ and $\mathcal{P}_a$, to minimize $\mathcal{L}(\mathcal{P}, \mathcal{P}_a)$, s.t. $\sum_{q_i} c(q_i) \leq b$.




(a) $G_0$ (b) Result of Q1 (c) Result of Q2

**Fig. 1.** Results of the two kinds of queries. (a) The adjacency matrix of $G_0$ where darker circles represent edges with higher weight. (b) The new edge weights $w'_{ij}$ after issuing the queries from Q1. (c) The graph expanded after issuing queries from Q2. The upper left corner of the matrix corresponds to $G_0$ and the remaining rows and columns correspond to the nodes in $V_1$.

## 2 Example Application and Related Work

The problem described above is inspired by our work in author coreference. Here we are given a set of citations that mention similar author names, and must partition them by the true identity of the author. As in our previous work [2], we build a graph in which nodes represent author mentions. The edge weights indicate the strength of our belief that two mentions refer to the same real author, and are estimated by a binary logistic regression classifier that uses features such as title, co-author overlap, etc. Note that, each partition should represent the set of mentions that correspond to the same real author.

Experimentally, we have shown significant accuracy improvement by making queries of type Q1 and Q2. In our case, we issue the queries to the web. We incorporate the results of the queries either as additional features or as additional nodes in the graph. For example, we can form a query by joining the titles of two citations and issuing it to a search engine API. A hit indicates the presence of a document on the web that contains both of these citations and hence provides some evidence that they are authored by the same person. The result of the query is translated into a binary input feature to our classifier and is used to update the weight on the corresponding edge. The problem is resource bounded because for a fully connected graph, obtaining additional feature value for every pair of mentions is prohibitively expensive.

Similarly, we can add nodes corresponding to documents obtained by web queries. Note that these web documents represent author mentions and help improve accuracy by transitivity. For example, the additional node could be the list of publications or CV of one of the authors and would show strong affinity towards several nodes in the original graph. Hence, by transitivity, applying graph partitioning on this expanded graph leads to improvement in accuracy. However, since the web is too large to incorporate all its data, we need an efficient procedure for selecting a subset of web queries and resulting documents.

In [2], we propose an approach to resource bounded information gathering based on expected entropy, in which we use web information as an additional feature. We also propose centroid-based methods in which we add nodes to the graph.

Learning and inference under resource limitations has been studied in various forms, including resource-bounded reasoning and the value of information [5], budgeted learning, [4], and active learning, for example, [3].

## 3 Acknowledgments

We thank Avrim Blum, Katrina Ligett, Chris Pal, Sridhar Mahadevan, Arnold Rosenberg, Gideon Mann, Siddharth Srivastava and Aron Culotta for useful discussions. Supported in part by the CIIR, CIA, NSA and NSF under grant #IIS-0326249 and in part by DoD contract #HM1582-06-1-2013.

## References

1. N. Bansal, S. Chawla, A. Blum, Correlation Clustering, Proc. of 43rd FOCS, 2002
2. P. Kanani, A. McCallum, C. Pal, Improving Author Coreference by Resource-bounded Information Gathering from the Web, Proc. of IJCAI, 2007
3. N. Roy, A. McCallum, Toward Optimal Active Learning through Sampling Estimation of Error Reduction, Proc. of 18th ICML, 2001
4. A. Kapoor, R. Greiner, Learning and Classifying Under Hard Budgets, ECML, 2005
5. J. Grass, S. Zilberstein, A Value-Driven System for Autonomous Information Gathering, JIIS, vol. 14, 2000