
Optimizing Large Scale Correlation Clustering

Shai Bagon Meirav Galun

Dept. of Computer Science and Applied Mathematics

The Weizmann Institute of Science

Rehovot 76100, Israel

<http://www.wisdom.weizmann.ac.il/~{bagon, meirav}>

Abstract

Clustering is a fundamental task in unsupervised learning. The focus of this paper is the Correlation Clustering functional which combines positive and negative affinities between the data points. The contribution of this paper is two fold: (i) Provide a theoretic analysis of the functional. (ii) New optimization algorithms which can cope with large scale problems ($> 100K$ variables) that are infeasible using existing methods. Our theoretic analysis provides a probabilistic generative interpretation for the functional, and justifies its intrinsic “model-selection” capability. Furthermore, we draw an analogy between optimizing this functional and the well known Potts energy minimization. This analogy allows us to suggest several new optimization algorithms, which exploit the intrinsic “model-selection” capability of the functional to automatically recover the underlying number of clusters. We compare our algorithms to existing methods on both synthetic and real data. In addition we suggest two new applications that are made possible by our algorithms: unsupervised face identification and interactive multi-object segmentation by rough boundary delineation.

1 Introduction

One of the fundamental tasks in unsupervised learning is clustering: grouping data points into coherent clusters. In clustering of data points, two aspects of pair-wise affinities can be measured: (i) *Attraction* (positive affinities), i.e., how likely are points i and j to be in the same cluster, and (ii) *Repulsion* (negative affinities), i.e., how likely are points i and j to be in different clusters.

Indeed, new approaches for clustering, recently presented by Yu and Shi [2001] and Bansal et al. [], suggest to combine attraction and repulsion information. Normalized cuts

was extended by Yu and Shi [2001] to allow for negative affinities. However, the resulting functional provides sub-optimal clustering results in the sense that it may lead to fragmentation of large homogeneous clusters.

The Correlation Clustering functional (CC), proposed by Bansal et al. [], tries to maximize the intra-cluster agreement (attraction) and the inter-cluster disagreement (repulsion). Contrary to many clustering objectives, the CC functional has an inherent “model-selection” property allowing to *automatically* recover the underlying number of clusters [Demaine and Immorlica].

Optimizing CC is tightly related to many graph partitioning formulations [Nowozin and Jegelka 2009], however it is known to be NP-hard [Bansal et al.]. Existing methods derive convex continuous relaxations to approximately optimize the CC functional. However, these algorithms do not scale beyond a few hundreds of variables. See for example, the works of [Nowozin and Jegelka 2009; Bagon et al. 2010; Vitaladevuni and Basri 2010; Glasner et al. 2011].

This work suggests a new perspective on the CC functional, showing its analogy to the known *Potts model*. This new perspective allows us to leverage on recent advances in discrete optimization to propose new CC optimization algorithms. We show that our algorithms scale to large number of variables ($> 100K$), and in fact can tackle tasks that were **infeasible in the past**, e.g., applying CC to pixel-level image segmentation. In addition, we provide a *rigorous statistical interpretation* for the CC functional and justify its intrinsic model selection capability. Our algorithms exploit this “model selection” property to automatically recover the underlying number of clusters k .

The contributions of this paper are as follows:

- A rigorous probabilistic interpretation of the CC functional, justifying its intrinsic model selection capability.
- A new perspective to the functional, drawing analogy to the discrete Potts model.
- New large scale optimization algorithms, that stem from our new perspective.
- Our algorithms automatically recover the underlying

number of clusters k .

- New applications in vision and graphics.

The first part of the paper (Sec. 2) focuses on the theoretical probabilistic interpretation of the CC functional. The subsequent sections are dedicated to the second part of this work which concerns the optimization of the CC functional.

Correlation Clustering (CC) Functional

Let $W \in \mathbb{R}^{n \times n}$ be an affinity matrix combining attraction and repulsion: for $W_{ij} > 0$ we say that i and j attract each other with certainty $|W_{ij}|$, and for $W_{ij} < 0$ we say that i and j repel each other with certainty $|W_{ij}|$. Thus the sign of W_{ij} tells us if the points attract or repel each other and the magnitude of W_{ij} indicates our certainty.

Any k -way partition of n points can be written as $U \in \{0, 1\}^{n \times k}$ s.t. $U_{ic} = 1$ iff point i belongs to cluster c . $\sum_c U_{ic} = 1 \forall i$ ensure that every i belongs to *exactly* one cluster.

The CC functional maximizes the intra-cluster agreement [Bansal et al.]. Given a matrix W^1 , an optimal partition U minimizes:

$$\begin{aligned} \mathcal{E}_{CC}(U) &= - \sum_{ij} W_{ij} \sum_c U_{ic} U_{jc} & (1) \\ \text{s.t. } U_{ic} &\in \{0, 1\}, \sum_c U_{ic} = 1 \end{aligned}$$

Note that $\sum_c U_{ic} U_{jc}$ equals 1 iff i and j belong to the same cluster. For brevity, we will denote $\sum_c U_{ic} U_{jc}$ by $[UU^T]_{ij}$ from here on.

2 Probabilistic Interpretation

This section provides a probabilistic interpretation for the CC functional. This interpretation allows us to provide a theoretic justification for the ‘‘model selection’’ property of the CC functional. Moreover, our analysis exposes the underlying implicit prior that this functional assumes.

We consider the following probabilistic generative model for matrix W . Let U be the true unobserved partition of n points into clusters. Assume that for some pairs of points i, j we observe their pairwise similarity values s_{ij} . These values are random realizations from either a distribution f^+ or f^- , depending on whether points i, j are in the same cluster or not. Namely,

$$\begin{aligned} p\left(s_{ij} = s \mid [UU^T]_{ij} = 1\right) &= f^+(s) \\ p\left(s_{ij} = s \mid [UU^T]_{ij} = 0\right) &= f^-(s) \end{aligned}$$

¹Note that W may be sparse. The ‘‘missing’’ entries are simply assigned ‘‘zero certainty’’ and therefore they do not affect the optimization.

Assuming independency of the pairs, the likelihood of observing similarities $\{s_{ij}\}$ given a partition U is then

$$\mathcal{L}(\{s_{ij}\} | U) = \prod_{ij} f^+(s_{ij})^{[UU^T]_{ij}} \cdot f^-(s_{ij})^{(1-[UU^T]_{ij})}$$

To infer a partition U using this generative model we look at the posterior distribution:

$$Pr(U | \{s_{ij}\}) \propto \mathcal{L}(\{s_{ij}\} | U) \cdot Pr(U)$$

where $Pr(U)$ is a prior. Assuming a *uniform prior* over all partitions, i.e., $Pr(U) = \text{const}$, yields:

$$Pr(U | \{s_{ij}\}) \propto \prod_{ij} f^+(s_{ij})^{[UU^T]_{ij}} \cdot f^-(s_{ij})^{(1-[UU^T]_{ij})}$$

Then, the negative logarithm of the posterior is given by

$$\begin{aligned} -\log Pr(U | \{s_{ij}\}) &= \hat{C} + \sum_{ij} \log f^+(s_{ij}) [UU^T]_{ij} \\ &\quad + \sum_{ij} \log f^-(s_{ij}) (1 - [UU^T]_{ij}) \end{aligned}$$

where \hat{C} is a constant not depending on U .

Interpreting the affinities as log odds ratios $W_{ij} = \log\left(\frac{f^+(s_{ij})}{f^-(s_{ij})}\right)$, the posterior becomes

$$-\log Pr(U | \{s_{ij}\}) = C - \sum_{ij} W_{ij} [UU^T]_{ij} \quad (2)$$

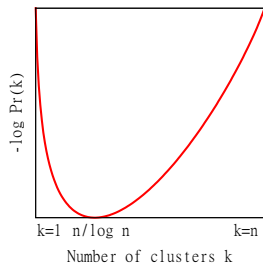
That is, Eq. (2) estimates the log-posterior of a partition U . Therefore, a partition U that minimizes Eq. (2) is the **MAP** (maximum a-posteriori) partition. Since Eq. (1) and Eq. (2) differ only by a constant they share the same minimizer: the MAP partition.

2.1 Recovering k (a.k.a. ‘‘model selection’’)

We showed that the generative model underlying the CC functional has a *single* model for all partitions, regardless of k . Therefore, optimizing the CC functional one need not select between different generative models to decide on the optimal k . Comparing partitions with different k is therefore straight forward and does not require an additional ‘‘model complexity’’ term (such as BIC, MDL, etc.)

As described in the previous section the CC functional assumes a uniform prior over all partitions. This uniform prior on U induces a prior on the number of clusters k , i.e., what is the a-priori probability of U having k clusters: $Pr(k) = Pr(U \text{ has } k \text{ clusters})$. We use Stirling numbers of the second kind [Rennie and Dobson 1969] to compute this induced prior on k . Fig 1 shows the non-trivial shape of this induced prior on the number of clusters k .

Figure 1: **Prior on the number of clusters k :** Graph shows $-\log \Pr(k)$, for uniformly distributed U . The induced prior on k takes a non-trivial shape: it assigns very low probability to the trivial solutions of $k = 1$ and $k = n$, while at the same time gives preference to partitions with non-trivial k . The mode of this prior is when U has roughly $\frac{n}{\log n}$ clusters.



3 CC Optimization: Continuous Perspective

After discussing the theory behind the CC functional and providing probabilistic justification to its model selection capability, we move to discuss methods and approaches to optimize this functional. We begin with a brief glance at current state-of-the-art CC optimization algorithms.

Optimizing the correlation clustering functional (Eq. (1)) is NP-hard [Bansal et al.]. Instead of solving **directly** for a partition U , existing methods optimize **indirectly** for the binary adjacency matrix $X = UU^T$, i.e., $X_{ij} = 1$ iff i and j belong to the same cluster. By introducing the binary adjacency matrix the quadratic objective (w.r.t. U): $-\sum_{ij} W_{ij} [UU^T]_{ij}$ becomes linear (w.r.t. X): $-\sum_{ij} W_{ij} X_{ij}$. The connected components of X , after proper rounding, are the resulting clusters, and the number of clusters k naturally emerges. Indirect optimization methods must ascertain that the feasible set consists only of “decomposable” X : $X = UU^T$. This may be achieved either by posing semi-definite constraints on X [Vitaladevuni and Basri 2010], or by introducing large number of linear inequalities [Demaine and Immorlica ; Vitaladevuni and Basri 2010]. These methods take a continuous and convex relaxation approach to approximate the resulting functional. This approach allows for nice theoretical properties due to the convex optimization at the cost of a very restricted scalability.

Solving for X requires $\sim n^2$ variables instead of only $\sim n$ when solving directly for U . Therefore, these methods scale poorly with the number of variables n , and in fact, they cannot handle more than a few hundreds of variables. In summary, these methods suffers from two drawbacks: (i) recovering U from X is highly susceptible to noise and more importantly (ii) it is *infeasible* to solve large scale problems by these methods.

4 Our New Perspective on CC

Existing methods view the CC optimization in the context of convex relaxation and build upon methods and ap-

proaches that are common practice in this field of continuous optimization. We propose an alternative perspective to the CC optimization: *viewing it as a discrete energy minimization*. This new perspective allows us to build upon recent advances in discrete optimization and propose efficient and direct CC optimization algorithms. More importantly, the resulting algorithms solve *directly* for U , and thus scales significantly better with the number of variables.

We now show how to cast the CC functional of Eq.(1) as a discrete pair-wise conditional random field (CRF) energy. For ease of notation, we describe a partition U using a labeling vector $L \in \{1, 2, \dots\}^n$: $l_i = c$ iff $U_{ic} = 1$. A general form of pair-wise CRF energy is $E(L) = \sum_i E_i(l_i) + \sum_{ij} E_{ij}(l_i, l_j)$ [Boykov et al. 2002]. Discarding the unary term ($\sum_i E_i(l_i)$), and taking the pair-wise term to be W_{ij} if $l_i \neq l_j$ we can re-write the CC functional as a CRF energy:

$$\mathcal{E}_{CC}(L) = \sum_{ij} W_{ij} \cdot \mathbb{1}_{[l_i \neq l_j]} \quad (3)$$

This is a Potts model. Optimizing the CC functional can now be interpreted as searching for a MAP assignment for the energy (3).

The resulting Potts energy has three unique characteristics, each posing a challenge to the optimization process:

- (i) **Non sub-modular:** The energy is non sub-modular. The notion of sub-modularity is the discrete analogue of convexity from continuous optimization [Lovasz 1983]. Optimizing a non sub-modular energy is NP-hard, even for the binary case [Rother et al. 2007].
- (ii) **Unknown number of labels:** Most CRF energies are defined for a fixed and known number of labels. Thus, the search space is restricted to $L \in \{1, \dots, k\}^n$. When the number of labels k is unknown the search space is by far larger and more complicated.
- (iii) **No unary term:** There is no unary term in the energy. The unary term plays an important role in guiding the optimization process [Szeliski et al. 2008]. Moreover, a strong unary term is crucial when the energy in non sub-modular [Rother et al. 2007].

There exist examples of CRFs in the literature that share some of these characteristics (e.g., non sub-modular [Rother et al. 2007; Kolmogorov and Wainwright 2005], unknown number of labels [Isack and Boykov 2011; Bleyer et al. 2010]). Yet, to the best of our knowledge, no existing CRF exhibits all these three challenges at once. More specifically, we are the first to handle non sub-modular energy that has no unary term. Therefore, we cannot just use “off-the-shelf” Potts optimization algorithms, but rather modify and improve them to cope with the three challenges posed by the CC energy.

Algorithm 1: Expand-and-Explore

Input: Affinity matrix $W \in \mathbb{R}^{n \times n}$
Output: Labeling vector $L \in \{1, 2, \dots\}^n$
Init $L_i \leftarrow 1, i = 1, \dots, n$ // initial labeling

repeat

 for $\alpha \leftarrow 1; \alpha \leq \#L + 1; \alpha ++$ **do**

 | $L \leftarrow \text{Expand}(\alpha)$
until L is unchanged

 $\#L$ denotes the number of different labels in L .

 $\text{Expand}(\alpha)$: expanding α using QPBOI.

 By letting $\alpha = \#L + 1$ the algorithm “expand” and explore an empty label. This may affect the number of labels $\#L$.

Algorithm 2: Swap-and-Explore

Input: Affinity matrix $W \in \mathbb{R}^{n \times n}$
Output: Labeling vector $L \in \{1, 2, \dots\}^n$
Init $L_i \leftarrow 1, i = 1, \dots, n$ // initial labeling

repeat

 for $\alpha \leftarrow 1; \alpha \leq \#L; \alpha ++$ **do**

 | for $\beta \leftarrow \alpha; \beta \leq \#L + 1; \beta ++$ **do**

 | | $L \leftarrow \text{Swap}(\alpha, \beta)$
until L is unchanged

 $\#L$ denotes the number of different labels in L .

 $\text{Swap}(\alpha, \beta)$: swapping labels α and β using QPBOI.

 By letting $\beta = \#L + 1$ the algorithm explore new number of clusters, this may affect the number of labels $\#L$.

5 Our Large Scale CC Optimization

In this section we adapt known discrete energy minimization algorithms to cope with the three challenges posed by the CC energy. We derive three CC optimization algorithms that stem from either large move making algorithms (α -expand and $\alpha\beta$ -swap [Boykov et al. 2002]), or Iterated Conditional Modes (ICM) [Besag 1986]. Our resulting algorithms scale gracefully with the number of variables n , and solve CC optimization problems that were infeasible in the past. Furthermore, our algorithms take advantage of the intrinsic model selection capability of the CC functional (Sec. 2) to robustly recover the underlying number of clusters.

5.1 Improved large move making algorithms

Boykov et al.[2002] introduced a very effective method for multi-label energy minimization that makes large search steps by iteratively solving binary sub-problems. There are two large move making algorithms: α -expand and $\alpha\beta$ -swap that differ by the binary sub-problem they solve. α -expand consider for each variable whether it is better to retain its current label or flip it to label α . The binary step of $\alpha\beta$ -swap involves only variables that are currently assigned to labels α or β , and consider whether it is better to retain their current label or switch to either α or β . Defined for sub-modular energies, the binary step in these algorithms is solved using graph-cut.

We propose new optimization algorithms: *Expand-and-Explore* and *Swap-and-Explore*, inspired by α -expand and $\alpha\beta$ -swap, that can cope with the challenges of the CC energy. (i) For the binary step we use a solver that handles non sub-modular energies. (ii) We incorporate “model selection” into the iterative search to recover the underlying number of clusters k . (iii) In the absence of unary term, a good initial labeling is provided to the non sub-modular binary solver.

Binary non sub-modular energies can be approximated by an extension of graph-cuts: QPBO [Rother et al. 2007]. When the binary energy is non sub-modular QPBO is not guaranteed to provide a labeling for all variables. Instead, it outputs only a partial labeling. How many variables are labeled depends on the amount of non sub-modular pairs and the relative strength of the unary term for the specific energy. When no unary term exists in the energy QPBO leaves most of the variables unlabeled. To circumvent this behavior we use the “improve” extension of QPBO (denoted by QPBOI): This extension is capable of improving an initial labeling to find a labeling with lower energy [Rother et al. 2007]. In the context of expand and swap algorithms a natural initial labeling for the binary steps is to use the current labels of the variables and use QPBOI to improve on it, ensuring the energy does not increase during iterations.

To overcome the problem of finding the number of clusters k our algorithms do not iterate over a fixed number of labels, but explore an “empty” cluster in addition to the existing clusters in the current solution. Exploring an extra empty cluster allows the algorithms to optimize over all solutions with any number of clusters k . The fact that there is no unary term in the energy makes it straight forward to perform. Alg. 1 and Alg. 2 presents our *Expand-and-Explore* and *Swap-and-Explore* algorithms in more detail.

5.2 Adaptive-label ICM

Another discrete energy minimization method that we modified to cope with the three challenges of the CC optimization is ICM [Besag 1986]. It is a point-wise greedy search algorithm. Iteratively, each variable is assigned the label that minimizes the energy, conditioned on the current labels of all the other variables. ICM is commonly used for MAP estimation of energies with a *fixed* number of labels. Here we present an *adaptive-label ICM*: using

the ICM conditional iterations we adaptively determine the number of labels k . Conditioned on the current labeling, we assign each point to the cluster it is most attracted to, or to a singleton cluster if it is repelled by all.

In this section we proposed a new perspective on CC optimization. Interpreting it as MAP estimation of Potts energy allows us to propose a variety of efficient optimization methods²:

- Swap-and-Explore (with binary step using QPBOI)
- Expand-and-Explore (with binary step using QPBOI)
- Adaptive-label ICM

Our proposed approach has the following advantages:

- It solves only for n integer variables. This is by far less than the number of variables required by existing methods described in Sec. 3, that requires $\sim n^2$ variables of the adjacency matrix $X = UU^T$. It makes our approach capable of dealing with large number of variables ($> 100K$) and suitable for pixel-level image segmentation.
- The algorithms solve directly for the cluster membership of each point, thus there is no need for rounding scheme to extract U from the adjacency matrix X .
- The number of clusters k is optimally determined by the algorithm and it does not have to be externally supplied like in many other clustering/segmentation methods.

In their work Elsner and Schudy [2009] proposed a greedy algorithm to optimize the CC functional over complete graphs. Their algorithm is in fact an ICM method presented outside the proper context of CRF energy minimization, and thus does not allow to generalize the concept of discrete optimization to more recent optimization methods.

6 Experimental Results

This section evaluates the performance of our proposed optimization algorithms using both synthetic and real data. We compare to both existing discrete optimization algorithms that can handle multi-label non sub-modular energies (TRW-S [Kolmogorov and Wainwright 2005] and BP [Pearl 1988]³), and to existing state-of-the-art CC optimization method of Vitaladevuni and Basri [2010]. Since existing CC optimization methods do not scale beyond several hundreds of variables, extremely small matrices are used in the following experiments. We leave it to Sec. 7 to evaluate our method on large scale problems.

²Matlab implementation available at: <http://www.wisdom.weizmann.ac.il/~bagon/matlab.html>.

³Since these two algorithms work only with pre-defined number of clusters k , we over-estimate k and report only the number of *non empty* clusters in the solution.

6.1 Synthetic data

This experiment uses synthetic affinity matrices W to compare our algorithms to existing Potts optimization algorithms. The synthetic data have 750 variables randomly assigned to 15 clusters with different sizes (ratio between larger to smaller cluster: $\sim \times 5$). For each variable we sampled roughly the same number of neighbors: of which $\sim 25\%$ are from within the cluster and the rest from the other clusters. We corrupted the clean ground-truth adjacency matrix with 20% noise affecting both the sign of W_{ij} and the certainty (i.e., $|W_{ij}|$). Overall the resulting percent of positive (sub-modular) connections is $\sim 30\%$.

We report several measurements for these experiments: run-time, energy (\mathcal{E}_{CC}), purity of the resulting clusters and the recovered number of clusters k for each of the algorithms as a function of the sparsity of the matrix W , i.e., percent of non-zero entries. Each experiment was repeated 10 times with different randomly generated matrices.

Fig. 2 shows results of the synthetic experiments. Existing multi-label approaches (TRW-S and BP) do not perform too well: higher \mathcal{E}_{CC} , lower purity and incorrect recovery of k . This demonstrates the difficulty of the energy minimization problem that has no unary term and many non sub-modular pair-wise terms. These results are in accordance with the observations of Kolmogorov and Wainwright [2005] when the energy is hard to optimize.

For our large move making algorithms, Expand-and-Explore provides marginally better clustering results than the Swap-and-Explore. However, its relatively slow running time makes it infeasible for large CC problems⁴. A somewhat surprising result of these experiments shows that for matrices not too sparse (above 10%), adaptive-label ICM performs surprisingly well. In fact, it is significantly faster than all the other methods and manages to converge to the correct number of clusters with high purity and low energy.

From these experiments we conclude that Swap-and-Explore (Alg. 2) is a very good choice of optimization algorithm for the CC functional. However, when the affinity matrix W is not too sparse, it is worth while giving our adaptive-label ICM a shot.

6.2 Co-clustering data

The following experiment compares our algorithms with a state-of-the-art CC optimization method, R-LP, of Vitaladevuni and Basri [2010]. For this comparison we use affinity matrices computed for co-segmentation. The co-

⁴This difference in run time between Expand and Swap can be explained by looking at the number of variables involved in each of the binary steps carried out: For the expand algorithm, each binary step involves almost all the variables, while the binary swap move considers only a small subset of the variables.

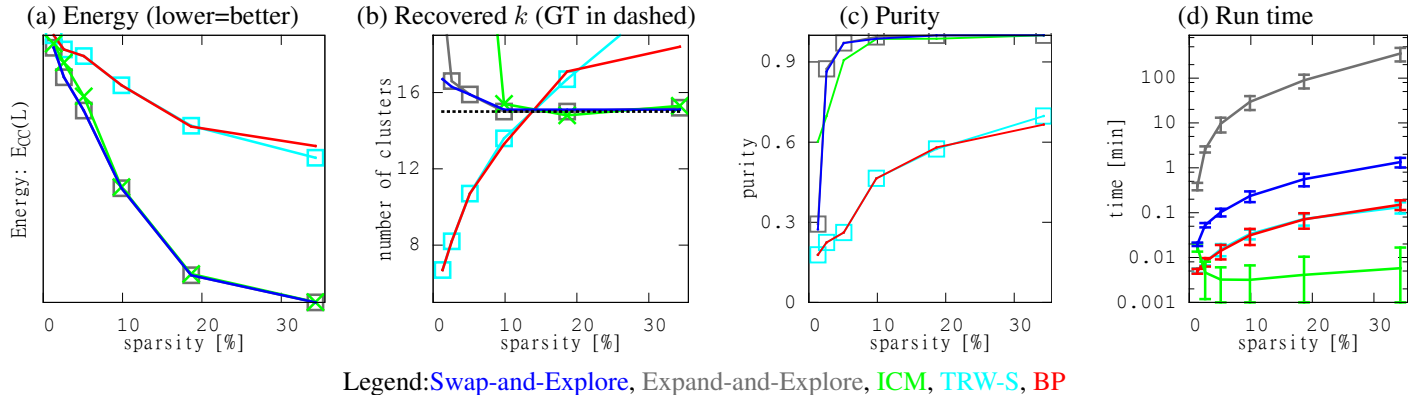


Figure 2: **Synthetic results:** Graphs comparing (a) Energy at convergence. (b) Recovered number of clusters. (c) Purity of resulting clusters. (d) Run time of algorithms (log scale). TRW-S and BP are almost indistinguishable, as are Swap and Expand in most of the plots.

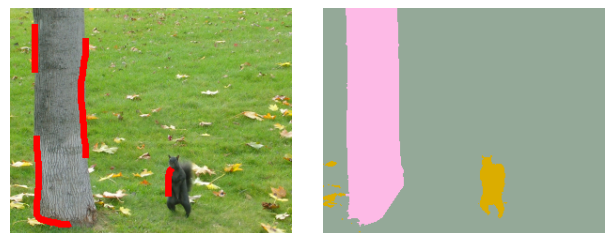
	Swap	Ours Expand	ICM	TRWS	BP
Energy ratio (%)	98.6 ±1.4	98.4 ±1.9	77.4 ±23.9	83.8 ±5.4	83.6 ±6.3
Strictly lower (> 100%)	15%	11.7%	0	0	0

Table 1: **Comparison to Glasner et al. [2011]:** Ratio between our energy and of Glasner et al.: Since all energies are negative, higher ratio means lower energy. Ratio higher than 100% means our energy is better than Glasner et al.. Bottom row shows the percentage of cases where each method got strictly lower energy than Glasner et al..

segmentation problem can be formulated as a correlation clustering problem with super pixels as the variables [Glasner et al. 2011].

We obtained 77 affinity matrices, courtesy of Glasner et al.[2011], used in their experiments. The number of variables in each matrix ranges from 87 to 788, Their sparsity (percent of non-zero entries) ranges from 6% to 50%, and there are roughly the same number of positive (sub-modular) and negative (non sub-modular) entries.

Table 1 shows the ratio between our energy and the energy of R-LP method. The table also shows the percent of matrices for which our algorithms found a solution with lower energy than R-LP. The results show the superiority of our algorithms to existing multi-label energy minimization approaches (TRW-S and BP). Furthermore, it is shown that our methods are in par with existing state-of-the-art CC optimization method on small problems. However, unlike existing methods, our algorithms can be applied to problems *two orders of magnitude larger*. Optimizing directly for U not only did not compromise the performance of our method, but also allows us to handle large scale CC optimization, as demonstrated in the next section.



(a) Input image and boundary scribbles (red)
(b) Resulting segmentation

Figure 3: **Interactive multi-object segmentation:** (a) The user provides only crude and partial indications to the locations of boundaries between the relevant objects in an image (red). (b) The output of our algorithm correctly segments the image into multiple segments. Image was taken from [Alpert et al. 2007].

7 New Applications

In this section we present two new applications made possible by our large scale CC optimization. Both these applications build upon integrates attraction and repulsion information between large number of points, and requires the robust recovery of the underlying number of clusters k .

7.1 Interactive multi-object segmentation (Patent Pending)

Our first experiment demonstrates the ability of our algorithm to handle large scale CC problem (pixel-level segmentation).

Negative affinities in image segmentation may come very naturally from boundary information: pixels on the same side of a boundary are likely to be in the same segment (attraction), while pixels on opposite sides of a boundary are likely to be in different segments (repulsion). We use this observation to design a novel approach to interactive multi-object image segmentation. Instead of using

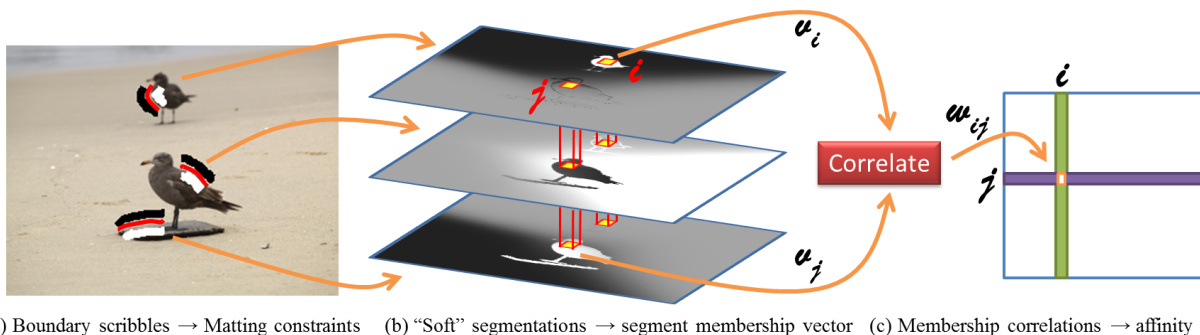


Figure 4: **From boundary scribbles to affinity matrix:** (a) A boundary scribble is drawn by the user (red), inducing “figure/ground” regions on its opposite sides (black and white regions). (b) For each scribble we use the method of Levine et al. [2008] to generate a soft segmentation of the image into two segments: pixel values in the soft segmentation are in the range $[-1, 1]$. Pixels far away from the scribble are assigned 0 as it is uncertain to what segment they should belong to. Each pixel i is described using a segmentation membership vector v_i with an entry corresponding to its assignment at each soft segmentation (red columns). (c) A non-zero entry w_{ij} in the sparse affinity matrix is the correlation between normalized vectors v_i and v_j : $w_{ij} = v_i^T v_j / \|v_i\| \cdot \|v_j\|$. We also add strong repulsion across each scribble.

k different “strokes” for the different objects (e.g., Santner et al. [2011]), the user applies a *single* “brush” to indicate parts of the boundaries between the different objects. Using these *sparse and incomplete* boundary hints we can correctly complete the boundaries and extract the desired number of segments. Although the user does not provide at any stage the number of objects k , our method is able to automatically detect the number of segments using only the *incomplete* boundary cues. Fig. 3 provides an example of our novel interactive multi-object segmentation approach.

Computing affinities: Fig. 4 illustrates how we use sporadic user-provided boundary cues to compute a *sparse* affinity matrix with both positive and negative entries. Note that this is a modification of the affinity computation presented by Stein et al. [2008]: (i) We use the interactive boundary cues to drive the computation, rather than some boundaries computed by unsupervised technique. (ii) We only compute a small fraction of all entries of the matrix, as opposed to the full matrix of Stein et al. (iii) Most importantly, we end up with both positive and negative affinities in contrast to Stein et al. who use only positive affinities.

The sparse affinity matrix W is very large ($\sim 100k \times 100k$). Existing methods for optimizing the correlation clustering functional are unable to handle this size of a matrix. We applied our Swap-and-Explore algorithm (Alg. 2) to this problem and it provides good looking results with only several minutes of processing per image.

Fig. 5 shows input images and user marked boundary cues used for computing the affinity matrix. Our results are shown at the bottom row.

The new interface allows the user to segment the image into several coherent segments without changing brushes and without explicitly enumerate the number of desired seg-

ments to the algorithm.

7.2 Clustering and face identification

Our next experiment is to show that detecting the underlying number of clusters k can be an important task on its own. Given a collection of face images we expect the different clusters to correspond to different persons. Identifying the different people requires not only high purity of the resulting clusters but more importantly to *correctly discover the appropriate number of clusters*. This experiment is an extension of existing work on the problem of “same/not-same” learning. Following recent metric learning approach (e.g., [Guillaumin et al. 2009; Guillaumin et al. 2010]) we learn a *single* classifier that assigns a probability for each pair of faces: “how likely is this pair to be of the same person”. Then, using this classifier, we are able to determine *the number of persons* and cluster the faces of *unseen people*. That is, given a new set of face images of several *unseen* people, our clustering approach is able to automatically cluster and identify how many different people are in the new set of face images of *never seen before* people.

For this experiment we use PUT face dataset [Kasinski et al. 2008]. The dataset consists of 9971 images of 100 people (roughly 100 images per person). Images were taken in partially controlled illumination conditions over a uniform background. The main sources of face appearance variations are changes in head pose, and facial expression.

We use the same method as Guillaumin et al. [2009] to describe each face. SIFT descriptors are computed at fixed points on the face at multiple scales. We use the annotations provided in the dataset to generate these keypoints. Given a training set of labeled faces $\{x_i, y_i\}_{i=1}^N$ we use a

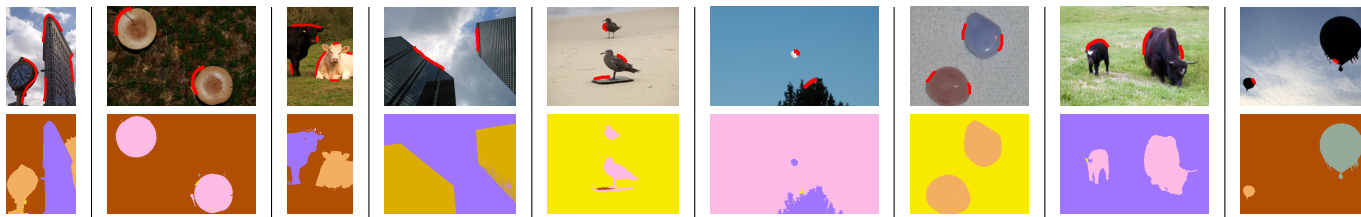


Figure 5: **Interactive segmentation results.** *Input image and user boundary cues (top), our result (bottom). Images were taken from [Alpert et al. 2007].*

state-of-the-art method by Guillaumin et al. [2010] to learn a Mahalanobis distance L and threshold b such that:

$$Pr(y_i = y_j | x_i, x_j; L, b) = \sigma \left(b - (x_i - x_j)^T L^T L (x_i - x_j) \right)$$

where $\sigma(z) = (1 - e^{-z})^{-1}$ is the sigmoid function.

For each experiment we chose k people for test (roughly $100 \cdot k$ images), and used the images of the other $100 - k$ people for training. The learned distance is then used to compute p_{ij} , the probability that faces i and j belong to the same person, for all pairs of face images of the k people in the test set. The affinities are set to $W_{ij} = \log \frac{p_{ij}}{1-p_{ij}}$. We apply our clustering algorithm to search for an optimal partition, and report the identified number of people k' and the purity of the resulting clusters. We experimented with $k = 15, 20, \dots, 35$. For each k we repeated the experiments for several different choices of k different persons.

In these settings all our algorithms performed roughly the same in terms of recovering k and the purity of the resulting clustering. However, in terms of running time adaptive-label ICM completed the task significantly faster than other methods. We compare Swap-and-Explore to two different approaches: (i) *Connected components*: Looking at the matrix of probabilities p_{ij} , thresholding it induces k' connected components. Each such component should correspond to a different person. At each experiment we tried 10 threshold values and reported the best result. (ii) *Spectral gap*: Treating the probabilities matrix as a *positive* affinity matrix we use NCuts [Shi and Malik 2000] to cluster the faces. For this method the number of clusters k' is determined according to the spectral gap: Let λ_i be the i^{th} largest eigenvalue of the normalized Laplacian matrix, the number of clusters is then $k' = \arg \max_i \frac{\lambda_i}{\lambda_{i+1}}$.

Fig. 6 shows cluster purity and the number of different persons k' identified as a function of the actual number of people k for the different methods. Our method succeeds to identify roughly the correct number of people (dashed black line) for all sizes of test sets, and maintain relatively high purity values.

8 Conclusion

This work provides generative probabilistic interpretation for the Correlation Clustering functional, justifying its in-

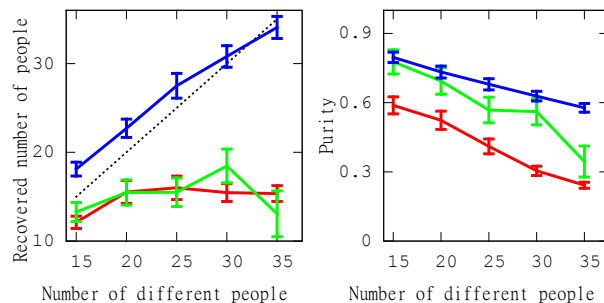


Figure 6: **Face identification:** *Graphs showing our result (Swap), spectral and connected components. Left: recovered number of people (k') vs. number of people in the test set. Dashed line shows the true number of people. Right: purity of resulting clusters.*

trinsic “model selection” capability. Using a generative probabilistic formulation allows for a better understanding of the functional, underlying assumptions it makes, including the prior it imposes on the solution.

Apart from establishing theoretic aspects of the CC functional, this work also suggests a new perspective on the functional, viewing it as a discrete Potts energy. The resulting energy minimization presents three challenges: (i) the energy is non sub-modular, (ii) the number of clusters is not known in advance, and (iii) there is no unary term. We proposed new energy minimization algorithms that can successfully cope with these challenges.

Optimizing large scale CC and robustly recovering the underlying number of clusters allows us to propose new applications: interactive multi-label image segmentation and unsupervised face identification.

Acknowledgments

The authors would like to thank these people for their fruitful and insightful remarks: Ronen Basri, Michal Irani, Boaz Nadler, Shiv Vitaladevuni, Daniel Glasner, Stella Yu, Tal Hassner and Lena Gorelick.

References

- Alpert, S., Galun, M., Basri, R., and Brandt, A. 2007. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *CVPR*.
- Bagon, S., Brostovsky, O., Galun, M., and Irani, M. 2010. Detecting and sketching the common. In *CVPR*.
- Bansal, N., Blum, A., and Chawla, S. Correlation clustering. *Machine Learning*.
- Besag, J. 1986. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*.
- Bleyer, M., Rother, C., and Kohli, P. 2010. Surface stereo with soft segmentation. In *CVPR*.
- Boykov, Y., Veksler, O., and Zabih, R. 2002. Fast approximate energy minimization via graph cuts. *PAMI*.
- Demaine, E. and Immorlica, N. Correlation clustering with partial information. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*.
- Elsner, M. and Schudy, W. 2009. Bounding and comparing methods for correlation clustering beyond ILP. In *ILP NLP*.
- Glasner, D., Vitaladevuni, S., and Basri, R. 2011. Contour-based joint clustering of multiple segmentations. In *CVPR*.
- Guillaumin, M., Verbeek, J., and Schmid, C. 2009. Is that you? Metric Learning Approaches for Face Identification. In *ICCV*.
- Guillaumin, M., Verbeek, J., and Schmid, C. 2010. Multiple instance metric learning from automatically labeled bags of faces. In *ECCV*.
- Isack, H. and Boykov, Y. 2011. Energy-based geometric multi-model fitting. *IJCV*.
- Kasinski, A., Florek, A., and Schmidt, A. 2008. The put face database. *Image Processing and Communications*. v13, 59–64.
- Kolmogorov, V. and Wainwright, M. 2005. On the optimality of tree-reweighted max-product message passing. In *Uncertainty in Artificial Intelligence*.
- Levin, A., Acha, R., and Lischinski, D. 2008. Spectral matting. *PAMI* 30, 10, 1699–1712.
- Lovasz, L. 1983. Submodular functions and convexity. *Mathematical programming: the state of the art*.
- Nowozin, S. and Jegelka, S. 2009. Solution stability in linear programming relaxations: Graph partitioning and unsupervised learning. In *ICML*.
- Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.
- Rennie, B. and Dobson, A. 1969. On stirling numbers of the second kind. *Journal of Combinatorial Theory*.
- Rother, C., Kolmogorov, V., Lempitsky, V., and Szummer, M. 2007. Optimizing binary MRFs via extended roof duality. In *CVPR*.
- Santner, J., Pock, T., and Bischof, H. 2011. Interactive multi-label segmentation. In *ACCV*.
- Shi, J. and Malik, J. 2000. Normalized cuts and image segmentation. *PAMI*.
- Stein, A., Stepleton, T., and Hebert, M. 2008. Towards unsupervised whole-object segmentation: Combining automated matting with boundary detection. In *CVPR*.
- Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., and Rother, C. 2008. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *PAMI*.
- Vitaladevuni, S. and Basri, R. 2010. Co-clustering of Image Segments Using Convex Optimization Applied to EM Neuronal Reconstruction. In *CVPR*.
- Yu, S. X. and Shi, J. 2001. Understanding popout through repulsion. In *CVPR*.