

# Clustering by transitive propagation

Vijay Kumar\* and Dan Levy\*

Simons Center for Quantitative Biology

Cold Spring Harbor Laboratory

1 Bungtown Road,

Cold Spring Harbor, NY 11724.

June 4, 2015

## Abstract

We present a global optimization algorithm for clustering data given the ratio of likelihoods that each pair of data points is in the same cluster or in different clusters. To define a clustering solution in terms of pairwise relationships, a necessary and sufficient condition is that belonging to the same cluster satisfies transitivity. We define a global objective function based on pairwise likelihood ratios and a transitivity constraint over all triples, assigning an equal prior probability to all clustering solutions. We maximize the objective function by implementing max-sum message passing on the corresponding factor graph to arrive at an  $\mathcal{O}(N^3)$  algorithm. Lastly, we demonstrate an application inspired by mutational sequencing for decoding random binary words transmitted through a noisy channel.

---

\*Email: [vsreeniv@cshl.edu](mailto:vsreeniv@cshl.edu), [levy@cshl.edu](mailto:levy@cshl.edu). Both authors contributed equally to this work.

# 1 Introduction

Most algorithms for clustering data points determine clusters by minimizing in-cluster differences. In this paper, we consider the clustering problem wherein the data points are governed by two likelihood functions:  $f_0(i, j)$  describing the probability that two data points  $i$  and  $j$  are from the same cluster, and  $f_1(i, j)$  describing the probability that  $i$  and  $j$  derive from different clusters. We use these two functions to assign a non-zero likelihood to any legal clustering configuration. This likelihood function is a product of  $f_0$  and  $f_1$  terms over all pairs of data-points. We include with this likelihood a second term that constrains the pair-wise assignments of “same” or “different” such that same-ness is transitive: a necessary and sufficient condition for ensuring a legal clustering configuration. This constraint term, acting on all triples  $(i, j, k)$ , determines a uniform prior on the space of all distinct clustering solutions.

As in the case of affinity propagation [1], we first describe the factor graph [2] determined by our likelihood function, and use max-sum message passing [3] to identify a clustering configuration that maximizes the posterior distribution given our observed data points. The result is a clustering algorithm that is  $\mathcal{O}(N^3 \times K)$  in complexity and  $\mathcal{O}(N^3)$  in memory usage, where  $N$  is the number of data-points and  $K$  is the number of iterations to convergence. In our experience, convergence is rapid and  $K$  is typically very small. The optimal clustering solution is a minimal energy configuration such that points are in the same cluster when they experience a net attractive force and in different clusters when the net force is repulsive. This algorithm has the added benefit of not requiring an *a priori* number of clusters.

In the next section, we calculate the posterior distribution whose maximization determines the optimal clustering. In section 3, we describe the factor graph for this distribution and describe our algorithm based on message passing. In section 4, we consider a detailed example that illustrates the method, and in section 5, we conclude with a summary of results, some trivial extensions, and future directions in applying relational constraints in factor graphs.

## 2 Calculating the posterior distribution

### Notation

Throughout this paper we will use the following notation.

$$\begin{aligned}
I &= \{1, 2, \dots, N\}, \text{ the data points,} \\
E &= \{(i, j) \mid i, j \in I, i \neq j, (i, j) \equiv (j, i)\}, \text{ the edges,} \\
T &= \{(i, j, k) \mid (i, j) \in E, (j, k) \in E, (k, i) \in E\}, \text{ the triples,} \\
f_0 &= P(i, j \mid i \text{ and } j \text{ are in the same cluster}) \\
f_1 &= P(i, j \mid i \text{ and } j \text{ are in different clusters})
\end{aligned}$$

We consider the fully connected graph  $G$  with nodes  $I$  and edges  $E$ . We assign a color to the edges of  $G$  such that any edge is either blue = 0 or red = 1. The hypothesis matrix is a function  $H : E \rightarrow \{0, 1\}$ ,

$$H_{ij} = \begin{cases} 0, & i, j \text{ belong to the same cluster (blue edge)} \\ 1, & i, j \text{ belong to different clusters (red edge)} \end{cases}$$

For any hypothesis matrix we can compute the likelihood as

$$L(I, H) = P(I|H) = \prod_{(i,j) \in E} f_1(i, j)^{H_{ij}} f_0(i, j)^{1-H_{ij}} \quad (1)$$

We assume that every clustering is equally likely, equivalent to a uniform prior over all  $H$  obeying the transitivity condition,

$$P_{prior}(H) = \begin{cases} \frac{1}{B_N} & , \quad H \text{ represents a valid clustering} \\ 0 & , \quad \text{otherwise} \end{cases} \quad (2)$$

Here  $B_N$  is the  $N$ -th Bell number that counts the total number of partitions of  $N$  data points.  $H$  represents a valid clustering when every triple  $(i, j, k) \in T$  satisfies the transitivity condition. The valid configurations for a single triple are shown in Figure 1. We can therefore

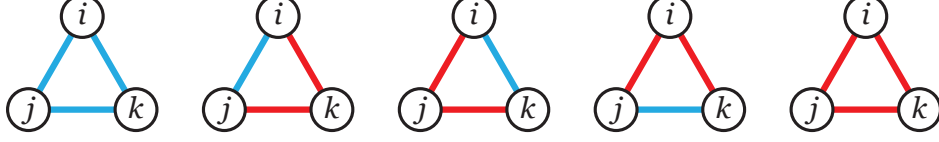


Figure 1: Valid configurations of hypothesis. Blue edge = 0 and red edge = 1.

express the uniform prior as a product over all triples:

$$P_{prior}(H) = \frac{1}{B_N} \prod_{(i,j,k) \in T} V_{ijk}, \text{ where} \quad (3)$$

$$V_{ijk} = \begin{cases} 1 & (H_{ij}, H_{jk}, H_{ki}) = (1, 1, 1), (1, 1, 0), (1, 0, 1), (0, 1, 1), (0, 0, 0) \\ 0 & (H_{ij}, H_{jk}, H_{ki}) = (0, 0, 1), (0, 1, 0), (1, 0, 0) \end{cases}. \quad (4)$$

For further details about the choice of prior and its consequences we refer the reader to Appendix A.

The posterior distribution over possible hypotheses  $H$  can be calculated using the likelihood function and prior defined above

$$P(H|I) = \frac{P(I|H)P_{prior}(H)}{\sum_H P(I|H)P_{prior}(H)}. \quad (5)$$

The sum in the denominator is the sum over all  $2^{N(N-1)/2}$  possible  $H$ . The prior restricts the posterior distribution to valid clustering solutions. We define the optimal clustering as the hypothesis matrix  $H^*$  that maximizes the posterior probability,

$$H^* = \operatorname{argmax}_H P(H|I). \quad (6)$$

$$= \operatorname{argmax}_H [\log P(I|H) + \log P_{prior}(H)] \quad (7)$$

$$= \operatorname{argmax}_H \left[ \sum_{(i,j) \in E} H_{ij} \log \frac{f_1(i,j)}{f_0(i,j)} + \sum_{(i,j,k) \in T} \log V_{ijk}(H_{ij}, H_{jk}, H_{ki}) \right]. \quad (8)$$

In arriving at the final result we have dropped terms that are independent of  $H_{ij}$  since they do not effect the result of the argmax operation.

To simplify notation we define an objective function

$$\boxed{\mathcal{F}(H) = \sum_{(i,j) \in E} S_{ij}(H_{ij}) + \sum_{(i,j,k) \in T} \delta_{ijk}(H_{ij}, H_{jk}, H_{ik}),} \quad (9)$$

where  $S_{ij}(H_{ij}) := H_{ij} \log \frac{f_1(i,j)}{f_0(i,j)}$  and  $\delta_{ijk} := \log V_{ijk}$ .

### Interpretation in terms of energy minimization

We can define a Hamiltonian or an energy function,

$$\mathcal{E} = - \sum_{(i,j) \in E} [H_{ij} \log f_1(i,j) + (1 - H_{ij}) \log f_0(i,j)] - \sum_{(i,j,k) \in T} \delta_{ijk}(H_{ij}, H_{jk}, H_{ki}) \quad (10)$$

over the space of all matrices  $H$ . Note that  $\mathcal{E} = -\mathcal{F} + \text{constant}$ . The optimal clustering is defined as the minimum of this energy function. The terms  $\log f_0$  and  $\log f_1$  can be viewed as forces of attraction and repulsion. For a given pair of points  $i, j$ , if  $f_0 > f_1$  then the energy is lowered if  $H_{ij} = 0$  or they are in the same cluster, and if  $f_1 > f_0$  the energy is lowered when  $H_{ij} = 1$ . In the absence of the prior term, the energy is minimized by the following solution

$$H_{ij}^{\text{no prior}} = \begin{cases} 0, & f_0(i,j) > f_1(i,j) \\ 1, & \text{otherwise} \end{cases}. \quad (11)$$

This solution is applicable when the data point clusters are well separated. Moreover, we have constructed this optimal solution through independent decisions for every edge. The prior complicates the problem and introduces a three-point long-range interaction term that is infinitely repulsive when the transitivity condition is disobeyed. However, if  $H_{ij}^{\text{no prior}}$  is consistent with transitivity, then it minimizes the energy and no further work is needed to identify an optimal configuration.

In the next section, we represent the objective function  $\mathcal{F}$  as a factor graph and use message passing to determine the configuration that maximizes the objective function.

## 3 Maximizing the objective function

We can represent the objective function and its dependence on the hypothesis matrix  $H$  with a *factor graph* [2]. The factor graph consists of two types of nodes: *variable nodes*, represented by a circle, for every independent hypothesis variable in  $H$ , and *function nodes*, represented by a square for each summand in the objective function (9). When a function node  $g$  depends on a variable  $x$ , we connect the nodes by an edge. Every variable node  $H_{ij}$  has  $(N - 2)$  edges that connect it to function nodes  $\delta_{ijk}$  for all  $k \neq i, j$ ; every function node

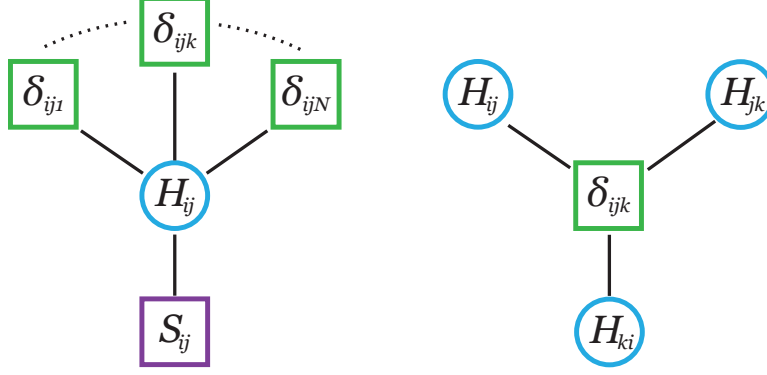


Figure 2: The factor graph for the objective function  $\mathcal{F}$  defined in equation (9) is composed of two types of junctions. On the left is the sub-graph of the neighbors of the variable node  $H_{ij}$  and on the right, the sub-graph of neighbors of the function node  $\delta_{ijk}$ .

$\delta_{ijk}$  is connected to three variable nodes  $H_{ij}$ ,  $H_{jk}$  and  $H_{ki}$ ; the function node  $S_{ij}$  has only one edge to the  $H_{ij}$  variable node. The factor graph is depicted in Figure 2.

We use message passing on the factor graph to solve for  $H^* = \operatorname{argmax}_H \mathcal{F}$ . This technique has been applied to a variety of problems in different fields as discussed in [4]. Since the factor graph has cycles, our approach is an example of loopy belief propagation [3]. The success of this method has been explained in terms of the accuracy of the Bethe free energy approximation [5]. Every message is a two-tuple as every hypothesis variable has two possible values. We denote the message transmitted from  $H_{ij}$  to  $\delta_{ijk}$  by  $\rho_{ij \rightarrow ijk}$  and the received message by  $\alpha_{ij \leftarrow ijk}$  as shown in Figure 3. Both messages are functions of the corresponding variable node  $H_{ij}$ . The function node  $S_{ij}$  continuously transmits the same message to  $H_{ij}$ .

The messages are updated as follows, first the variable nodes transmit to function nodes

$$\rho_{ij \rightarrow ijk}(H_{ij}) = S_{ij}(H_{ij}) + \sum_{l \neq i, j, k} \alpha_{ij \leftarrow ijl}(H_{ij}), \quad (12)$$

and then receive responses

$$\alpha_{ij \leftarrow ijk}(H_{ij}) = \max_{H_{jk}, H_{ki}} [\delta_{ijk}(H_{ij}, H_{jk}, H_{ki}) + \rho_{jk \rightarrow ijk}(H_{jk}) + \rho_{ki \rightarrow ijk}(H_{ki})]. \quad (13)$$

This sequence of transmission and reception defines one iteration of the algorithm. At the end of each iteration, the configuration  $H^*$  is given by

$$H_{ij}^* = \operatorname{argmax}_{x \in \{0,1\}} \left( S_{ij}(x) + \sum_{k \neq i, j} \alpha_{ij \leftarrow ijk}(x) \right) \quad (14)$$

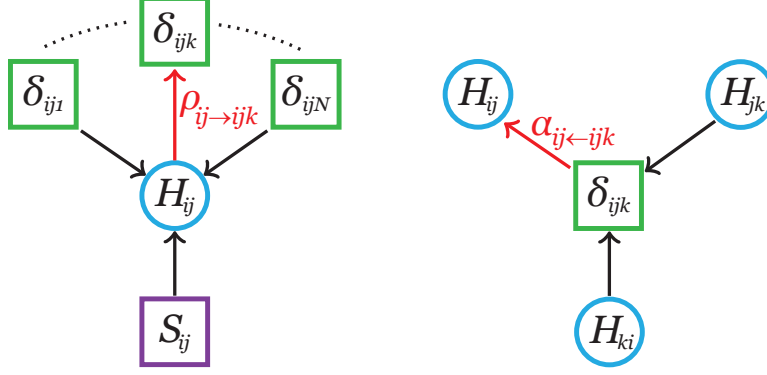


Figure 3: As in Figure 2, we show the transmitted and received messages for  $H_{ij}$ .

We repeat, iterating through transmissions and receptions until  $H^*$  is unchanged.

The message update rules can be considerably simplified. First, the  $\rho_{ij \rightarrow ijk}$  messages can be eliminated, such that we need only compute updates for  $\alpha_{ij \leftarrow ijk}$ . Second, the solution  $H_{ij}^*$  only depends on the combination  $A_{ijk} := \alpha_{ij \leftarrow ijk}(1) - \alpha_{ij \leftarrow ijk}(0)$  so we do not need to calculate values for both states (blue and red) but only for the difference. Lastly, we introduce the auxiliary matrix  $B_{ij} := \Delta S_{ij} + \sum_{k \neq i, j} A_{ijk}$  that reduces the complexity of the update procedure from  $\mathcal{O}(N^4)$  to  $\mathcal{O}(N^3)$ . We refer the interested reader to the discussion in Appendix B for details. Here, we show the result in the form of an explicit algorithm that we call *Transitive Propagation*, which has complexity  $\mathcal{O}(N^3 \times K)$  and  $\mathcal{O}(N^3)$  memory usage, where  $K$  is the (typically small) number of iterations to convergence.

---

**Algorithm:** Transitive Propagation

---

**Data:**  $N$  data-points with distributions  $f_{0,1}(i, j)$  for all  $i, j \in I$  and  $\lambda \in (0, 1)$ .

**Result:** Optimal hypothesis matrix  $H^*$ .

Calculate  $N \times N$  matrix  $\Delta S_{ij} = \log f_1(i, j) - \log f_0(i, j)$ .

Initialize  $N \times N \times N$  matrix  $A_{ijk} := 0$ ;

Define convergence goal  $M = 1000$ ;

**do**

    Compute  $N \times N$  matrix  $B$  :  $B_{ij} = \Delta S_{ij} + \sum_{k \neq i, j} A_{ijk}$  ;

    Compute update  $\Delta A_{ijk}$  defined by

$$\Delta A_{ijk} = \max \{0, B_{jk} - A_{jki} + B_{ki} - A_{kij}\} - \max \{0, B_{jk} - A_{jki}, B_{ki} - A_{kij}\} ;$$

    Perform update including the dampening factor  $\lambda$ :

$$A_{ijk} \leftarrow (1 - \lambda) A_{ijk} + \lambda \Delta A_{ijk} ;$$

    Calculate  $\Delta B_{ij} = \sum_{k \neq i, j} \Delta A_{ijk}$  and

$$m = - \min_{(i, j) \in E} \frac{B_{ij}}{\Delta B_{ij}} .$$

**while**  $0 < m < M$ ;

    Compute  $N \times N$  matrix  $H^*$ :  $H_{ij}^* = \begin{cases} 1, & B_{ij} \geq 0 \\ 0, & B_{ij} < 0 \end{cases}$

---

### Convergence and dampening

We have introduced a dampening factor  $\lambda$  that helps the algorithm converge to a fixed point rather than a cycle. Small values of  $\lambda$  promote convergence but also increase the running time of the algorithm. We find that the choice  $\lambda = 0.5$  is a good balance between time to convergence and avoiding cycles.

The entries in  $A_{ijk}$  do not converge to fixed values, and this is to be expected because we do not normalize the messages after each iteration. The solution  $\{H_{ij}^*\}$  only depends on the sign of the  $B_{ij}$  matrix. Consequently, our convergence criterion is as follows: at each iteration we estimate the minimum number of iterations,  $m$ , it would take to change the sign of one entry in  $B_{ij}$  and stop when the number of iterations reaches a defined threshold,  $M$ .



## 4 Example: clustering random bit patterns

In this section, we present the clustering problem that inspired the development of transitive propagation. Recently, one of the authors proposed a method to uniquely tag DNA molecules through a process of random mutagenesis. By marking each template molecule with a random pattern, we can resolve two difficulties that continue to plague high-throughput short-read sequencing: (1) counting DNA molecules accurately and (2) assembling DNA sequences across repeat regions that exceed a read length. We do not discuss the details here, but refer the reader to the original paper [6].

The example we address in this section is an abstracted version of the first problem, known in the literature as the  $K$ -populations problem, and has been shown to be NP-hard [7]. Assume we have  $K$  initial copies of a DNA sequence containing  $L$  mutable positions. Our mutation protocol randomly assigns one of two letters with equal probability at each position, generating  $K$  binary words of length  $L$ . These templates are copied many times and a machine analyzes those copies, outputting a *read* that matches the initial template's binary word but introduces errors at a rate of  $p_e \ll 1$  per bit. Starting with  $N$  reads generated through this process, we would like to determine the number of initial templates  $K$ , assigning the reads to clusters that correspond to the same initial template.

We work in a regime where  $N \gg K$  so that all templates are sampled and read by the sequencer. Since the error rate is low, we expect that the  $N$  reads form  $K$  clusters, where  $K$  is the unknown number of templates that we wish to determine. We begin by measuring the hamming distance  $d_{ij}$  between all reads  $i$  and  $j$ . When two reads are in the same cluster,

$$f_0(i, j) = \binom{L}{d_{ij}} x^{d_{ij}} (1 - x)^{L - d_{ij}}, \text{ where } x = 2p_e(1 - p_e) \quad (15)$$

and when they belong to different clusters,

$$f_1(i, j) = \binom{L}{d_{ij}} \frac{1}{2^L}. \quad (16)$$

We generated  $K$  templates of length  $L = 30$  bits and generated  $N = 10K$  reads by uniform sampling. We introduced errors at a rate of  $p_e$  per bit. The results that we present were obtained by averaging over 100 simulations for various values of  $K$  and  $p_e$ .

We performed computer simulations to evaluate our algorithm. We generated  $K$  random templates of length  $L = 30$  bits for  $K = 10, 20, 40$ . We simulated  $N = 10K$  reads with various error rates of  $p_e = 0.01, 0.05, 0.10, 0.15$  per bit. Figure 4 shows the accuracy in determination

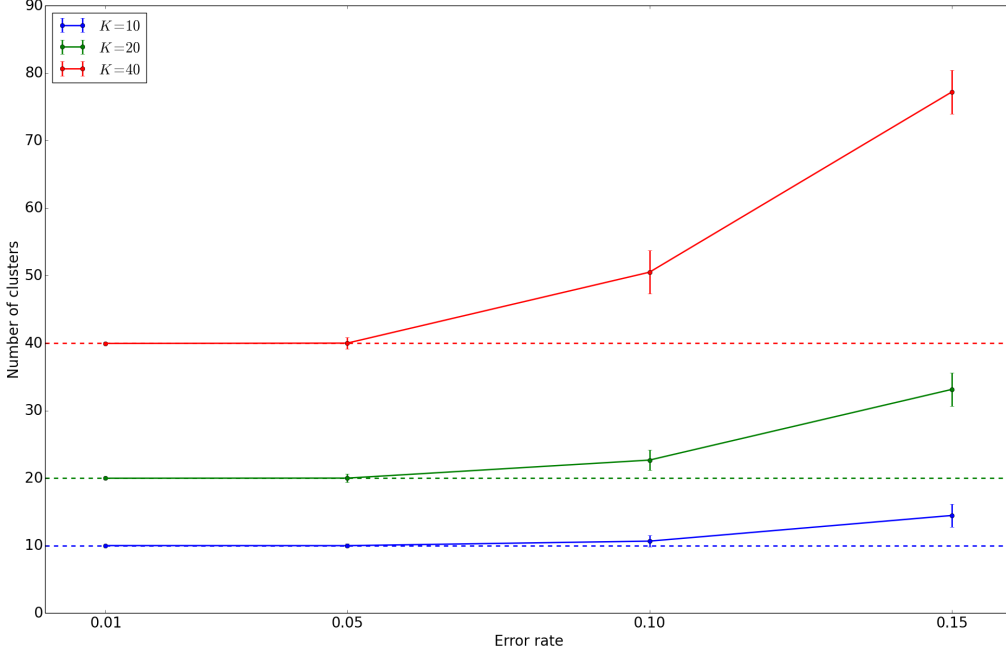


Figure 4: The number of clusters obtained as a function of the error rate for different template counts, which are shown as dashed lines. The results are averaged over 100 simulations and the error bars denote one standard deviation.

of the template count as a function of the error rate averaged over 100 simulations. We see accurate recovery of the template count even at high error rates of  $p_e = 0.05$ .

Our algorithm is also very accurate in determining the correct clustering configuration when the error rate is high. We fixed  $K = 50$  templates of length  $L = 30$  bits and generated  $N = 250$  reads for various values of error rate  $p_e = 0.01, 0.05, 0.10, 0.20$  and performed 100 simulations. Our measure of accuracy is the number of edges that are mis-classified by the algorithm averaged over all the simulations. We plot the number of incorrect edges as a function of the hamming distance between the reads in Figure 5. As a reference, we also plot the number of incorrect edges if we classified each edge  $i, j$  as red or blue based only on the likelihood ratio  $f_1(i, j)/f_0(i, j)$ . As expected, edges with very low or very high hamming distance are correctly inferred using both methods. For edges in the intermediate regime our method makes better inferences due to the transitive property.

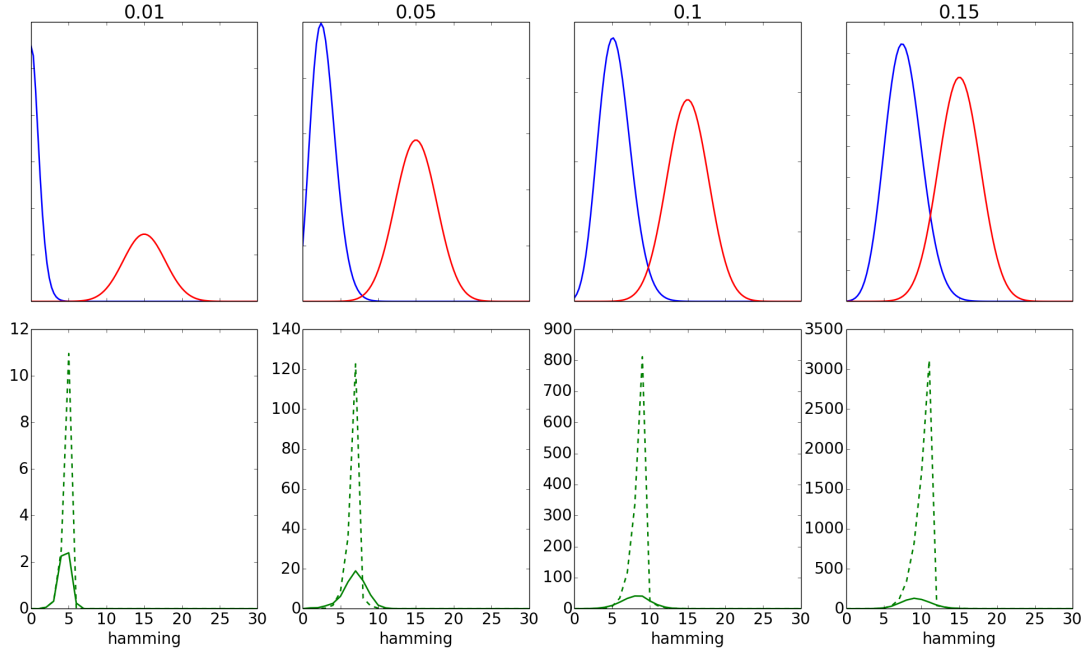


Figure 5: The panel above shows the distributions  $f_0$  (in blue) and  $f_1$  (in red) for various values of the error rate  $p_e$ . In the corresponding panel below we show the average number of incorrect calls made by our algorithm (green, solid line) and by classifying each edge  $i, j$  based on the likelihood ratio  $f_1(i, j)/f_0(i, j)$  alone.

## 5 Discussion

Transitive propagation is a useful algorithm for clustering data modeled by a balance of attractive and repulsive factors. By imposing a naive prior, the method uniformly explores the space of all partitions of the data-points, enforcing no *a priori* number of clusters or arbitrary similarity cut-off as required by other methods. As described in Appendix A, the naive prior does impose a non-uniform probability on the number of clusters. However, even this prior distribution may be tuned.

The transitive propagation algorithm can be extended in the following ways. First, the existing algorithm implements max-sum message passing to identify a single configuration that maximizes the likelihood. However, we can also implement sum-product message passing to determine the marginal posterior probabilities that two data-points derive from a common cluster. Such an algorithm would allow the selection of only the most confident edges, so as to discard outlying data-points. Second, the existing framework assumes that

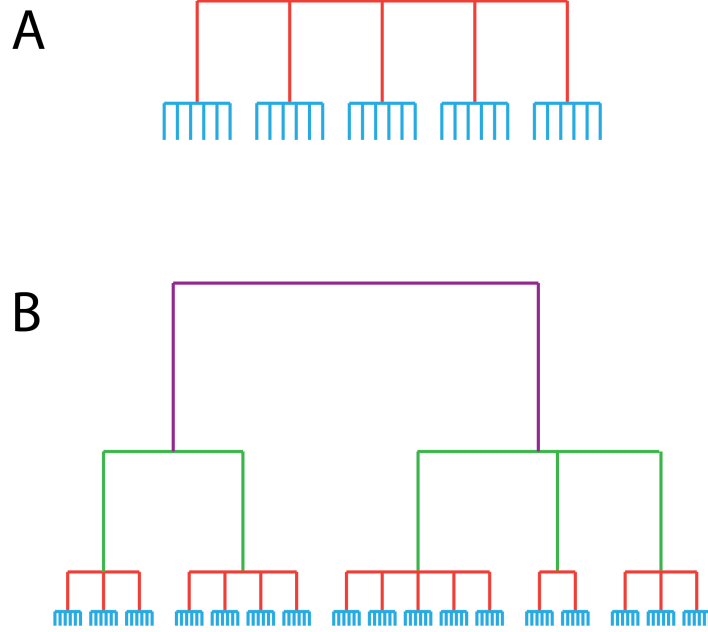


Figure 6: Panel A shows an ultrametric tree corresponding to the transitivity propagation algorithm. In panel B, we extend the ultrametric property to include four distinct levels, each with a likelihood function  $f_\alpha(i, j)$  for  $\alpha = 0, 1, 2, 3$  corresponding to multi-scale clustering.

the  $f$  function depends on  $i$  and  $j$  such that all clusters follow the same distribution. This limitation can be overcome through the inclusion of node-specific clustering parameters that enable variation in the intra-cluster distributions.

The methodology used in this paper to address clustering can be extended to other problems that we leave to future work. First, the transitive constraint may be considered as the first non-trivial example of an integer valued ultra-metric which assumes one of two values: 0 or 1. In this formulation, the prior constraint on  $V_{ijk}$  in equation 4 is identical to the ultra-metric property. We can extend to higher order clusters by allowing a family of likelihood functions,  $f_\alpha$  for  $\alpha = 0, 1, \dots, L$ , that measure increasingly divergent relationships between nodes (see figure 6) and allow  $H_{ij}$  to assume values of 0, 1, ...  $L$ . This modified algorithm enables multi-scale clustering. Second, we can apply the same framework of constrained optimization to enforce relationships other than equality. For example, we may have data-points that obey a partial ordering. The same constraints apply to  $H_{ij}$  as before, however it is no longer the case that  $H_{ij} = H_{ji}$ . Depending on the nature of the data, the optimization function may depend on the four possible states for the pair  $(H_{ij}, H_{ji})$  equivalent to the four

possible cases: (1)  $i$  and  $j$  are the coincident, (2)  $i$  precedes  $j$ , (3)  $j$  precedes  $i$ , or (4) there is no relation between  $i$  and  $j$ .

## Acknowledgements

We thank Robert Aboukhalil, Arjun Bansal, Sharat Chikkerur, Vishaka Datta, Sarah Harris, Ivan Iossifov, Jude Kendall, Bud Mishra, Swagatam Mukhopadhyay, Adam Siepel, Vinay Satish, Michael Schatz, Michael Wigler, Boris Yamrom, and the participants of QB Tea on May 13, 2015 for discussions, questions, and feedback that helped develop our ideas. VK and DL are funded by CSHL grant 125217/QB-SIMONS. This work was also supported by a grant from the Simons Foundation (SFARI award number 235988).

## References

- [1] B. J. Frey and D. Dueck, “Clustering by passing messages between data points,” *science*, vol. 315, no. 5814, pp. 972–976, 2007.
- [2] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, “Factor graphs and the sum-product algorithm,” *Information Theory, IEEE Transactions on*, vol. 47, no. 2, pp. 498–519, 2001.
- [3] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.
- [4] M. Mézard, “Passing messages between disciplines,” *Science*, vol. 301, no. 5640, pp. 1685–1686, 2003.
- [5] J. S. Yedidia, W. T. Freeman, and Y. Weiss, “Understanding belief propagation and its generalizations,” *Exploring artificial intelligence in the new millennium*, vol. 8, pp. 236–239, 2003.
- [6] D. Levy and M. Wigler, “Facilitated sequence counting and assembly by template mutagenesis,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 43, pp. E4632–E4637, 2014.
- [7] L. Parida and B. Mishra, “Partitioning single-molecule maps into multiple populations: algorithms and probabilistic analysis,” *Discrete applied mathematics*, vol. 104, no. 1, pp. 203–227, 2000.

## A Some observations about our choice of prior

We can construct a family of conjugate priors for the problem of clustering  $N$  data points parameterized by a real  $N \times N$  matrix  $X = [X_{ij}]$ ,

$$P_{conj}(H \mid X, N) = \frac{1}{\mathcal{Z}(X, N)} \prod_{(i,j) \in E} X_{ij}^{H_{ij}} \times \prod_{(i,j,k) \in T} V_{ijk}, \quad (17)$$

where  $\mathcal{Z}$  is a normalization factor. With this choice of prior, the posterior distribution in equation (5) becomes

$$P(H \mid I, X, N) = P_{conj}(H \mid X', N), \quad (18)$$

where  $X'_{ij} = X_{ij} + \log f_1(i, j) - \log f_0(i, j)$ . In this section, we study a one-parameter sub-family given by  $X_{ij} = x$ , where  $x$  is a non-negative real number,

$$F(H, x, N) = \frac{1}{Z(x, N)} \prod_{(i,j) \in E} x^{H_{ij}} \times \prod_{(i,j,k) \in T} V_{ijk}. \quad (19)$$

The uniform prior introduced in (3) is a member of this one-parameter family,  $P_{prior}(H) = F(H, x = 1, N)$ . The function  $Z(x, N)$  is the overall normalization and is usually called the partition function

$$Z(x, N) = \sum_H e^{-\mathcal{H}(H, x, N)} \quad (20)$$

where the Hamiltonian  $\mathcal{H}(H, x, N) = \sum_{(i,j,k) \in T} \delta_{ijk}(H_{ij}, H_{jk}, H_{ik}) + \sum_{(i,j) \in E} H_{ij} \log x$  is an example of a spin Hamiltonian where the  $H_{ij}$  can be viewed as “spin” degrees of freedom and the parameter  $\log x$  is the applied magnetic field. However, rather than the usual pairwise spin-spin interaction we have a 3-spin  $\delta_{ijk}$  term. We study the phase diagram of this Hamiltonian as a function of  $x$  and find an order-disorder transition at a critical value of  $x = x_{critical}$  where  $x_{critical} - 1 \sim \mathcal{O}(\frac{\log N}{N})$ . We suspect that this system has been studied in the vast literature on spin Hamiltonians and spin glasses but we are not aware of this.

Alternatively, it can be written as a sum over configurations satisfying the transitivity constraint

$$Z(x, N) = \sum_C x^{b(C)}, \quad (21)$$

where  $C$  runs over all possible partitions of  $N$  data points into clusters and  $b(C)$  is the number of blue edges. When  $x = 1$  we recover the uniform prior (3); and  $Z(x = 1, N) = B_N$  where  $B_N$  are the Bell numbers that enumerate the total number of partitions of a set of  $N$  elements. The limiting behavior is

$$\lim_{x \rightarrow 0} Z(x, N) = 1, \quad \lim_{x \rightarrow \infty} \frac{Z(x, N)}{x^{N(N-1)/2}} = 1. \quad (22)$$

The partition function  $Z(x, N)$  satisfies a recurrence relation

$$Z(x, N+1) = \sum_{k=0}^N \binom{N}{k} x^{k(k+1)/2} Z(x, N-k), \quad (23)$$

$$Z(x, 0) = Z(x, 1) = 1, \quad (24)$$

which can be derived using the principle of induction. This relation can be used to compute values of  $Z(x, N)$  numerically.

Intuitively, the effect of the parameter  $x$  is to favor or disfavor clustering configurations based on the number of blue edges. This can be quantified by calculating the number of blue edges averaged over the space of clustering configurations using the prior distribution (19)

$$\langle b \rangle = x \frac{d}{dx} \log Z(x, N). \quad (25)$$

The behavior of the blue edge fraction is shown in Figure 7. We see a phase transition, which in the  $N \rightarrow \infty$  limit is a discontinuity at  $x = 1$ . Phase transitions of this sort occur in the large  $N$  limit and arise when there is a balance between entropic and energetic considerations. When  $x = 1 + \epsilon$ ,  $\epsilon > 0$ , there is an exponentially larger weight associated with configurations with more blue edges. In the large  $N$  limit,  $\log Z(1 + \epsilon, N) = N(N-1)/2 \log(1 + \epsilon)$ , which is the contribution to the partition function from the configuration with all blue edges. This is the ordered phase. We estimate the entropy associated with the number of clustering configurations in the large  $N$  limit as  $\log Z(1, N) \sim N \log N$ . The balance gives us an estimate of the location of the phase transition as

$$x_{critical} = 1 + \epsilon, \quad \epsilon \sim \frac{2}{N} \log N. \quad (26)$$

The family of priors in (19) imposes a non-uniform prior on the number of clusters. To

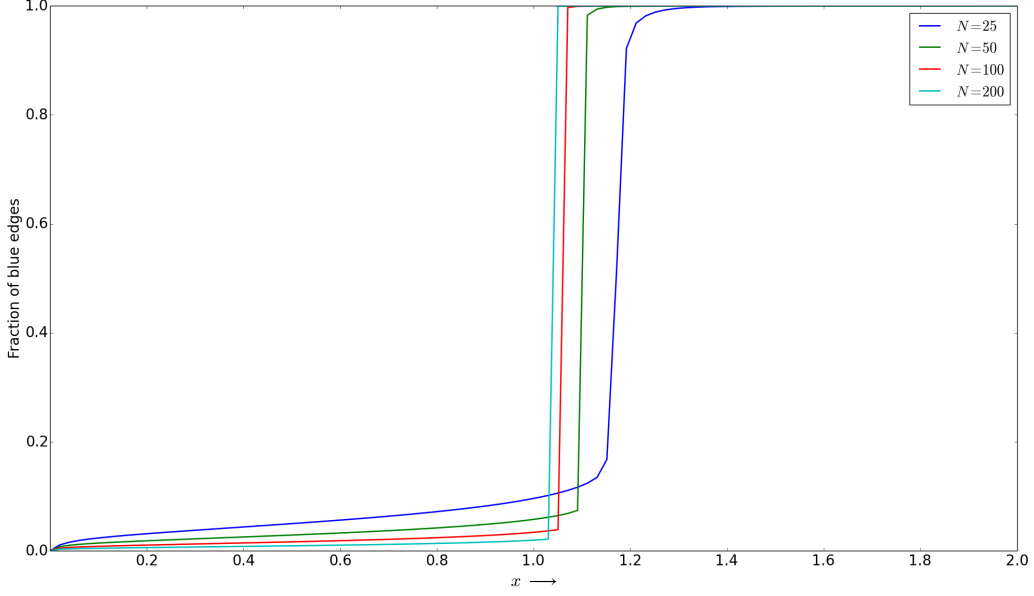


Figure 7: The expected value of the fraction of blue edges averaged over the ensemble of clustering configurations depicted as a function of the parameter  $x$  for various values of  $N$ .

calculate expectation values we add another parameter  $\lambda$  to the partition function

$$Z_\lambda(x, N) = \sum_C x^{b(C)} \lambda^{n(C)}, \quad (27)$$

where  $n(C)$  is the number of clusters, and the sum is over all clustering configurations. Clearly,  $Z_{\lambda=1}(x, N) = Z(x, N)$ , and taking derivatives with respect to  $\lambda$  allows us to calculate moments

$$\mu_n = \langle n \rangle = \left. \frac{d}{d\lambda} Z_\lambda(x, N) \right|_{\lambda=1}, \quad (28)$$

$$\sigma_n = \langle n^2 \rangle - \langle n \rangle^2 = \left. \left( \lambda \frac{d}{d\lambda} \right)^2 Z_\lambda(x, N) \right|_{\lambda=1}. \quad (29)$$



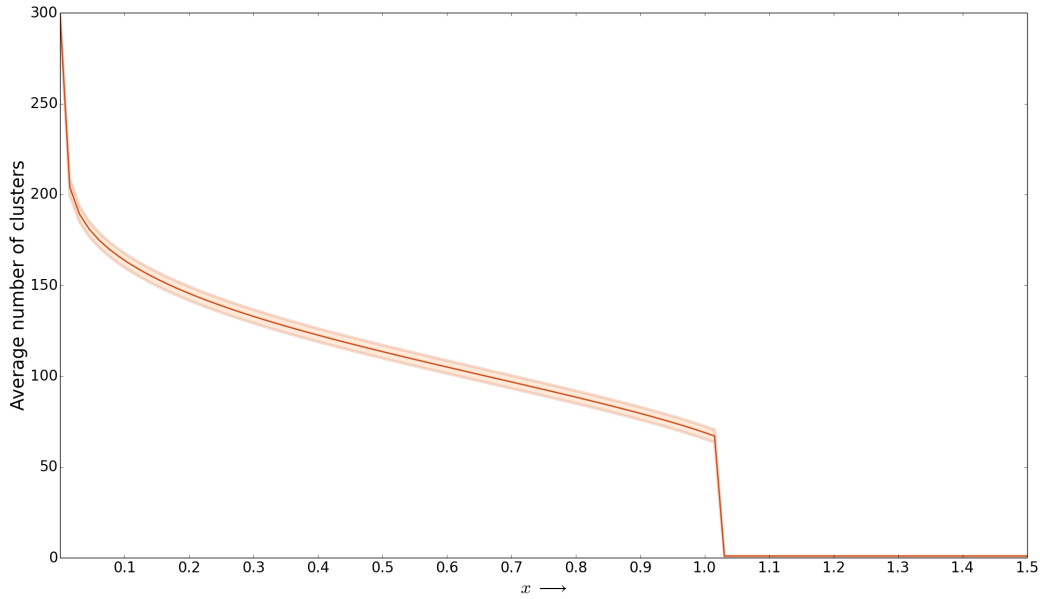


Figure 8: The line shows the mean number of clusters as a function of  $x$ . The shading indicates one standard deviation on either side. The plot is for  $N = 300$  points.

The function  $Z_\lambda(x, N)$  can be calculated using the recurrence relation

$$Z_\lambda(x, N + 1) = \lambda \sum_{k=0}^N \binom{N}{k} x^{k(k+1)/2} Z_\lambda(x, N - k), \quad (30)$$

$$Z_\lambda(x, 0) = 1, \quad Z_\lambda(x, 1) = \lambda. \quad (31)$$

The prior distribution is peaked over configurations with a definite number of clusters as shown in Figure 8.

We note that the parameter  $x$  can be tuned by the user between 0 and 1 in order to influence the outcome of the clustering based on prior knowledge either about the number of clusters or the fraction of blue edges. We recommend the uniform prior corresponding to the choice  $x = 1$  that weighs all clustering configurations equally.

## B Simplifying the message update equations

We recall here the message update equations from section 3.

$$\rho_{ij \rightarrow ijk}(H_{ij}) = S_{ij}(H_{ij}) + \sum_{l \neq i, j, k} \alpha_{ij \leftarrow ijl}(H_{ij}), \quad (32)$$

$$\alpha_{ij \leftarrow ijk}(h_{ij}) = \max_{H_{jk}, H_{ki}} [\delta_{ijk}(H_{ij}, H_{jk}, H_{ki}) + \rho_{jk \rightarrow ijk}(H_{jk}) + \rho_{ki \rightarrow ijk}(H_{ki})]. \quad (33)$$

### Simplification 1: eliminate $\rho$

Since the messages  $\rho_{ij \rightarrow ijk}$  play no role in determining the solution  $H^*$ , they can be eliminated giving a single update for the message  $\alpha_{ij \leftarrow ijk}$ , given below

$$\alpha_{ij \leftarrow ijk}(H_{ij}) = \max_{H_{jk}, H_{ki}} \left( \delta_{ijk} + S_{jk}(H_{jk}) + \sum_{l \neq i, j, k} \alpha_{jk \leftarrow jkl}(H_{jk}) + S_{ki}(H_{ki}) + \sum_{l \neq i, j, k} \alpha_{ki \leftarrow ikl}(H_{ki}) \right). \quad (34)$$

### Simplification 2: Only the difference of messages matters

Equation (14) can be rewritten as

$$H_{ij}^* = \text{Step} \left( S_{ij}(1) - S_{ij}(0) + \sum_{k \neq i, j} [\alpha_{ij \leftarrow ijk}(1) - \alpha_{ij \leftarrow ijk}(0)] \right), \quad (35)$$

$$\text{Step}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (36)$$

Note that  $H^*$  is only dependent on the differences

$$\Delta S_{ij} := S_{ij}(1) - S_{ij}(0), \quad A_{ijk} := \alpha_{ij \leftarrow ijk}(1) - \alpha_{ij \leftarrow ijk}(0). \quad (37)$$

Moreover, as we shall see shortly, the update for the difference  $A_{ijk}$  is completely determined by the value of  $A_{ijk}$  alone.

Equation (34) can be written explicitly as

$$\begin{aligned} \alpha_{ij \leftarrow ijk}(1) &= \max_{H_{jk}, H_{ki} = (1,1), (1,0), (0,1)} \left( S_{jk}(H_{jk}) + \sum_{l \neq i, j, k} \alpha_{jk \leftarrow jkl}(H_{jk}) + S_{ki}(H_{ki}) + \sum_{l \neq i, j, k} \alpha_{ki \leftarrow ikl}(H_{ki}) \right), \\ &= S_{jk}(0) + \sum_{l \neq i, j, k} \alpha_{jk \leftarrow jkl}(0) + S_{ki}(0) + \sum_{l \neq i, j, k} \alpha_{ki \leftarrow ikl}(0) \\ &\quad + \max \left\{ 0, \Delta S_{jk} + \sum_{l \neq i, j, k} A_{jkl}, \Delta S_{ki} + \sum_{l \neq i, j, k} A_{kil} \right\}. \end{aligned} \quad (38)$$

In the first step the  $\delta_{ijk}$  either contributes  $-\infty$  or nothing at all. Since the  $-\infty$  contribution never wins in the  $\max()$  function those configurations of  $H_{ij}, H_{jk}, H_{ki}$  are effectively eliminated. Similarly,

$$\begin{aligned}
\alpha_{ij \leftarrow ijk}(0) &= \max_{H_{jk}, H_{ki}=(1,1), (0,0)} \left( S_{jk}(H_{jk}) + \sum_{l \neq i, j, k} \alpha_{jk \leftarrow jkl}(H_{jk}) + S_{ki}(H_{ki}) + \sum_{l \neq i, j, k} \alpha_{ki \leftarrow ikl}(H_{ki}) \right), \\
&= S_{jk}(0) + \sum_{l \neq i, j, k} \alpha_{jk \leftarrow jkl}(0) + S_{ki}(0) + \sum_{l \neq i, j, k} \alpha_{ki \leftarrow ikl}(0) \\
&\quad + \max \left\{ 0, \Delta S_{jk} + \sum_{l \neq i, j, k} A_{jkl} + \Delta S_{ki} + \sum_{l \neq i, j, k} A_{kil} \right\}. \tag{39}
\end{aligned}$$

Taking the difference of equations (39), (38), we arrive at an update for the  $A_{ijk}$ :

$$\begin{aligned}
A_{ijk} \leftarrow & \max \left\{ 0, \Delta S_{jk} + \sum_{l \neq i, j, k} A_{jkl} + \Delta S_{ki} + \sum_{l \neq i, j, k} A_{kil} \right\} - \\
& \max \left\{ 0, \Delta S_{jk} + \sum_{l \neq i, j, k} A_{jkl}, \Delta S_{ki} + \sum_{l \neq i, j, k} A_{kil} \right\}.
\end{aligned}$$

In each iteration of the algorithm we have to update  $\mathcal{O}(N^3)$  variables  $A_{ijk}$ , each of which involves a sum over  $\mathcal{O}(N)$  terms. This makes the complexity  $\mathcal{O}(N^4)$ . The run time scaling with  $N$  can be improved to  $\mathcal{O}(N^3)$  by computing the summations ahead of time. We introduce the matrix  $B_{ij} := \Delta S_{ij} + \sum_{k \neq i, j} A_{ij}(k)$  in terms of which the update to the  $A_{ijk}$  becomes

$$A_{ijk} \leftarrow \max \{0, B_{jk} - A_{jki} + B_{ki} - A_{kij}\} - \max \{0, B_{jk} - A_{jki}, B_{ki} - A_{kij}\}, \tag{40}$$

and the best configuration is obtained by

$$H_{ij}^* = \text{Step}(B_{ij}). \tag{41}$$