Semi-supervised Clustering using Combinatorial MRFs

Ron Bekkerman,
University of Massachusetts

Joint work with Mehran Sahami (Google)

Multi-modal clustering

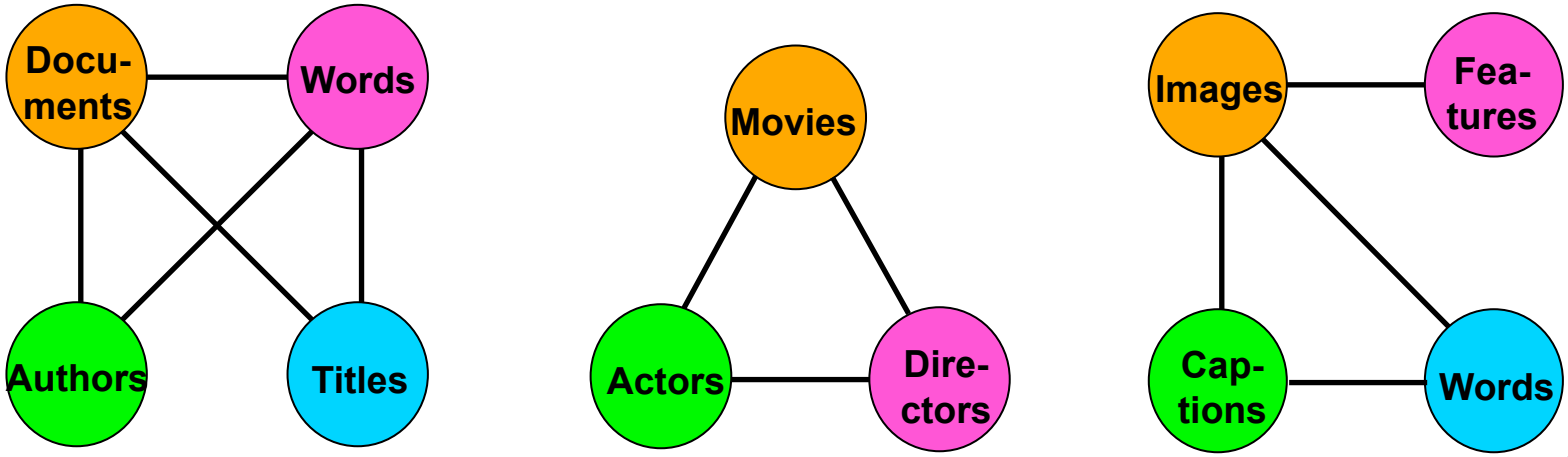
- Simultaneously constructing N clusterings of N modalities of the data
  - E.g. clustering documents, their words, authors, titles, markup elements, etc.
  - Clustering each modality helps clustering all the others
- Other multi-modal learning problems:
  - Multi-modal ranking
  - Multi-modal filtering

Last year we proposed

Bekkerman et al. ICML-2005

- A model for multi-modal clustering
  - where interactions between modalities are described using:

Pairwise Interaction Graph

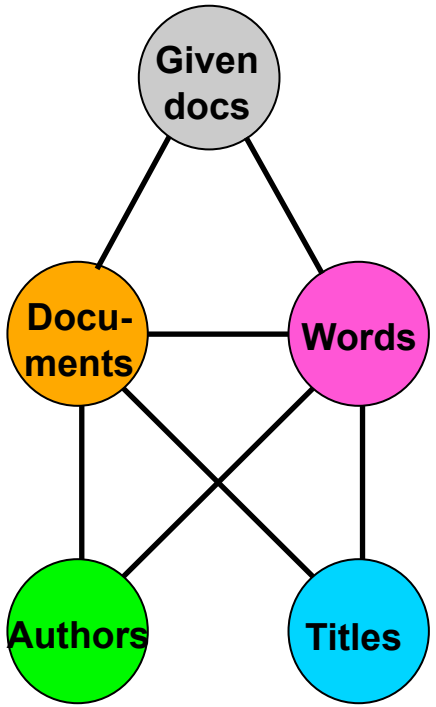


Proposed objective

- Let $(\tilde{X}, E)$ be pairwise interaction graph
- Extending *Dhillon et al. KDD-2003*:
- **Objective:** weighted sum of pairwise MI
  - $$\max_{\tilde{X}_1,\ldots,\tilde{X}_N} \sum_{(\tilde{X}_i,\tilde{X}_j)\in E} w_{ij}\, \mathrm{I}(\tilde{X}_i;\tilde{X}_j)$$
  - Subject to $\left|\tilde{X}_i\right| = K_i, \;\; i = 1,\ldots,N$
- No multi-dimensional probability tables
- Can be easily factorized

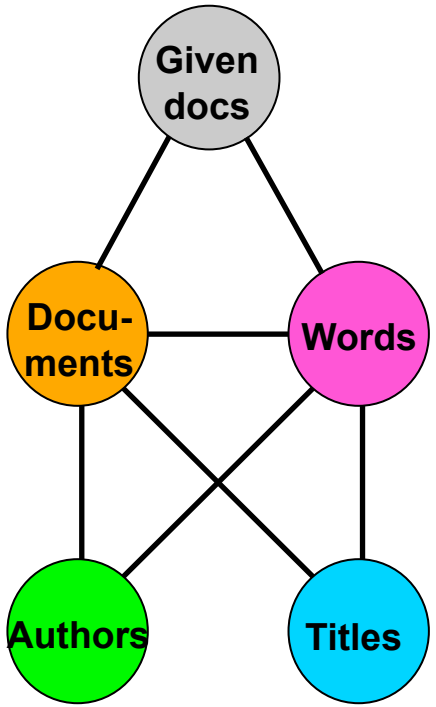
Semi-supervised case

- Natural generalization
- Fundamental problems:
  - Pairwise interaction graph has no probabilistic interpretation
  - *"Given docs"* is not a modality



Possible solution

- "Documents" is a random variable
  - Over all possible partitionings of dataset
- "Given docs" is an observed random variable
  - Whose value is a given partitioning



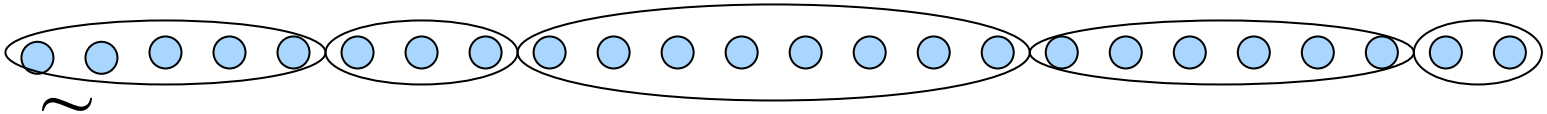
Combinatorial random variable

- Discrete random variable $\tilde{X}^c$ defined over a combinatorial set
  - Given a set X of n values
  - $\tilde{X}^c$ is defined over a set of $O(2^n)$ values
- Example: **lotto 6/49**
  - Given a set of 49 balls, draw 6 balls
  - $\tilde{X}^c$ is defined over *all* the subsets of size 6
  - $\binom{49}{6} = 13{,}983{,}816$ values

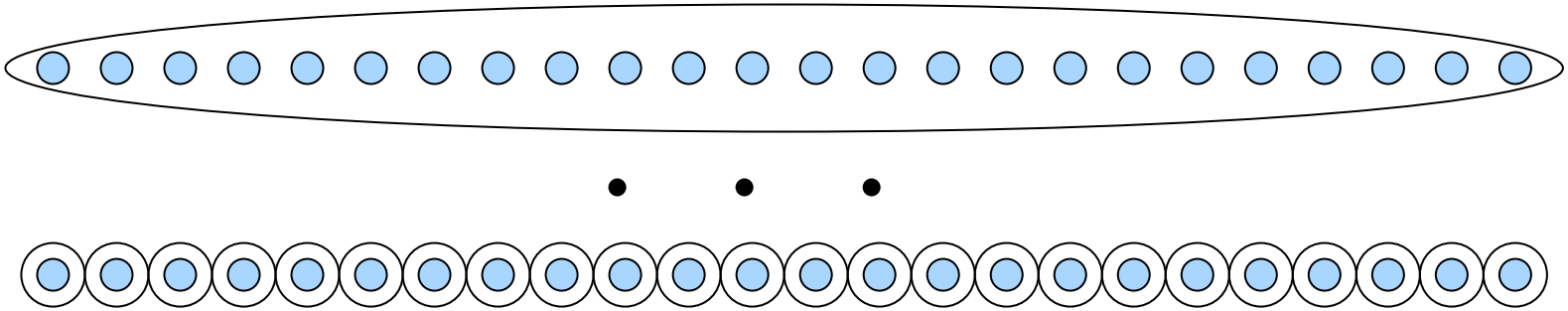
Example: hard clustering

- X is a r.v. over the data (n data points)

- $\tilde{X}$ is a r.v. over a partitioning of the data



- $\tilde{X}^c$ is a r.v. over all possible partitionings



- $O(k^n)$ values (k is number of clusters)

Combinatorial MRF (Comraf)

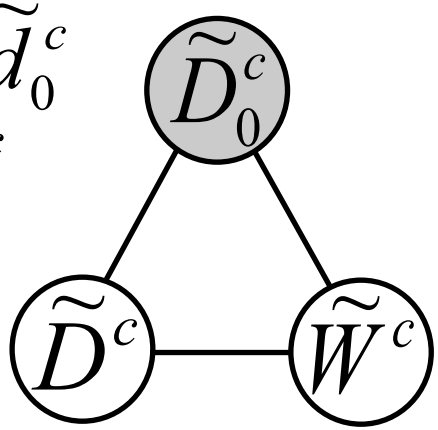
- MRF with combinatorial random variables
- Goal:
  - Find “best” (most likely) assignment to combinatorial random variables
- Comraf model: graph G and objective F
- Challenge:
  - Usually, $P(\tilde{X}^c)$ cannot be explicitly specified
  - No existing inference methods applicable

Properties of Comraf models

- **Neither generative nor discriminative**
  - No generative assumptions required
  - No arbitrarily chosen priors
- **Compact:** one node per “concept”
  - Such as *clusterings of documents*, *rankings of movies*, *subsets of images* etc.
  - Model learning is feasible
- **Generic:** applicable to many tasks
  - In unsupervised & semi-supervised learning

Semi-supervised clustering

- *Intrinsic* Comraf model
- We are given some labeled data of D
  - Which forms natural partitioning $\tilde{d}_0^c$
  - Represented as observed r.v. $\tilde{D}_0^c$
  - With a r.v. $\tilde{D}_0$ defined over $\tilde{d}_0^c$



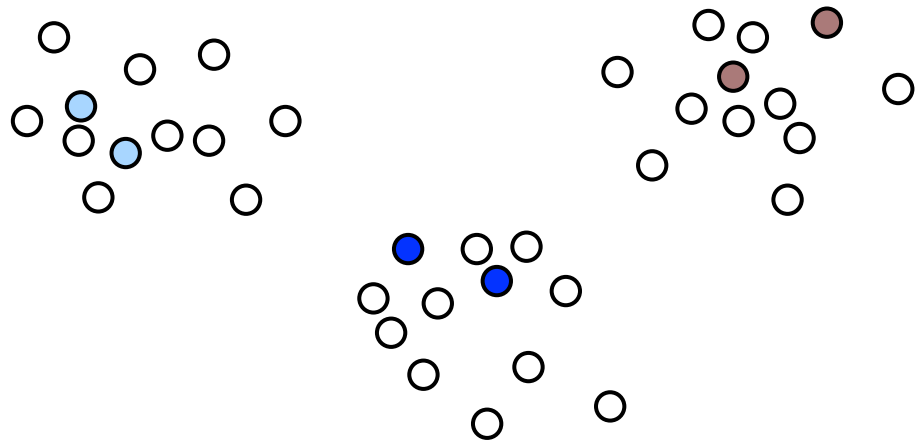
- Objective:

$$\max_{\tilde{d}^c, \tilde{w}^c} \mathrm{I}(\tilde{D};\tilde{W}) + \mathrm{I}(\tilde{D};\tilde{D}_0) + \mathrm{I}(\tilde{W};\tilde{D}_0)$$

- Algorithmic setup is the same

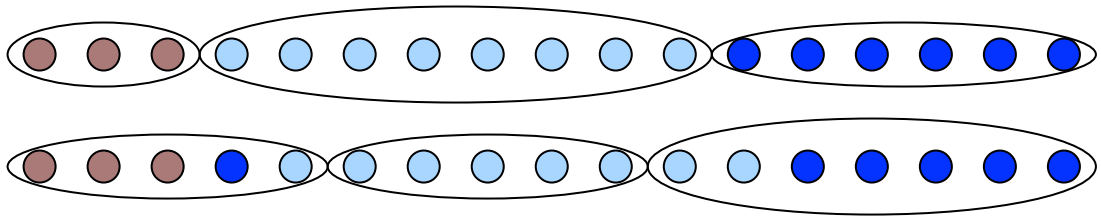
Constrained optimization scheme

- Well-established approach to semi-supervised clustering
  - *Wagstaff & Cardie ICML-2000* and others
- Must-link and cannot-link constraints



Evaluation methodology

- Clustering evaluation
  - Is generally unintuitive
  - Is an entire research field
- We use the “accuracy” measure
  - Following Slonim et al. and Dhillon et al.
  - Ground truth:
  - Our results:



- $$Acc = \frac{1}{|X|}\sum_{c}\gamma_c$$

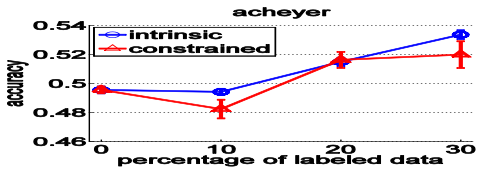
Size of dominant class in cluster c

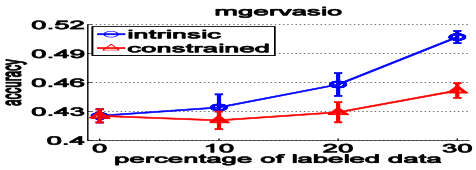
Datasets

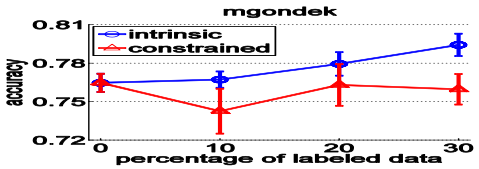
- Three CALO email datasets:
  - acheyer: 664 messages, 38 folders
  - mgervasio: 777 messages, 15 folders
  - mgondek: 297 messages, 14 folders
- Two Enron email datasets:
  - kitchen-l: 4015 messages, 47 folders
  - sanders-r: 1188 messages, 30 folders
- The 20 Newsgroups: 19,997 messages

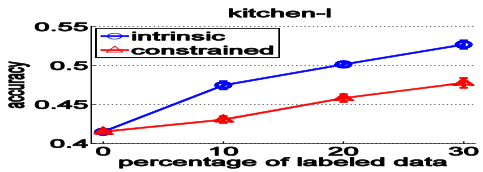
Results on email datasets

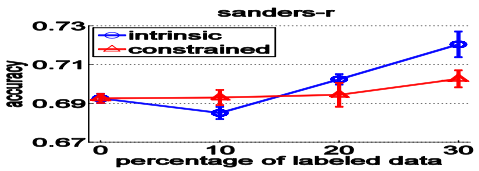
- Randomly choose 10, 20 and 30% of data to be labeled
- Plot the accuracy of the unlabeled portion









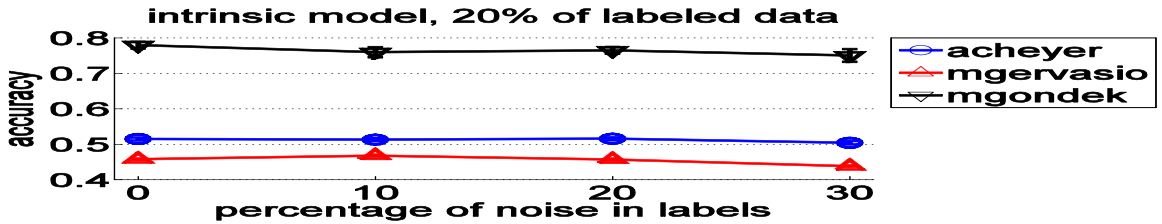


Semi-supervised clustering on 20NG

- 69.5±0.7% unsupervised clustering
- We consider 10% of data as labeled
- 74.8±0.6% constrained scheme
- 78.9±0.8% intrinsic Comraf scheme
- 3 of 5 runs built *well-balanced* clusterings
  - Where *each* category is dominant in one cluster
  - Clustering can be compared with classification
- 80.0±0.6% on 3 well-balanced clusterings
- 77.2±0.2% SVM on the same data

Resistance to noise

- Intrinsic scheme is resistant to noise
  - In contrast to constrained scheme
- Randomly corrupt 10, 20 and 30% labels:



Conclusion

- Comraf is a new type of graphical model
  - Useful at least for multi-modal clustering
  - But other applications are also developed
- The model is generic
  - Semi-supervised case is straightforward
- Inference algorithms are effective
  - And efficient (sub-cubic)
- Model learning is possible