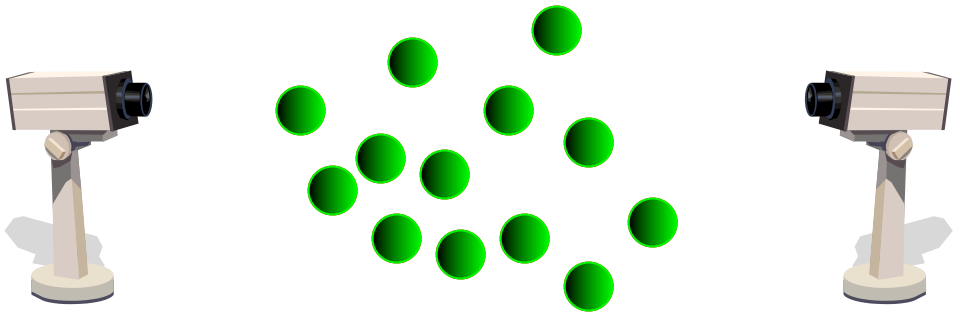
Combinatorial Markov Random Fields

Ron Bekkerman,
University of Massachusetts, USA

Joint work with Mehran Sahami (Google)
and Erik Learned-Miller (UMass)

Multi-modal learning

- Essential aspect of unsupervised learning
- Datasets usually have various views
  - Or various *modalities*
  - Such as: *documents*, *words*, *authors*, *titles* etc.



- Modalities shed light on *structure* of data

Multi-modal clustering

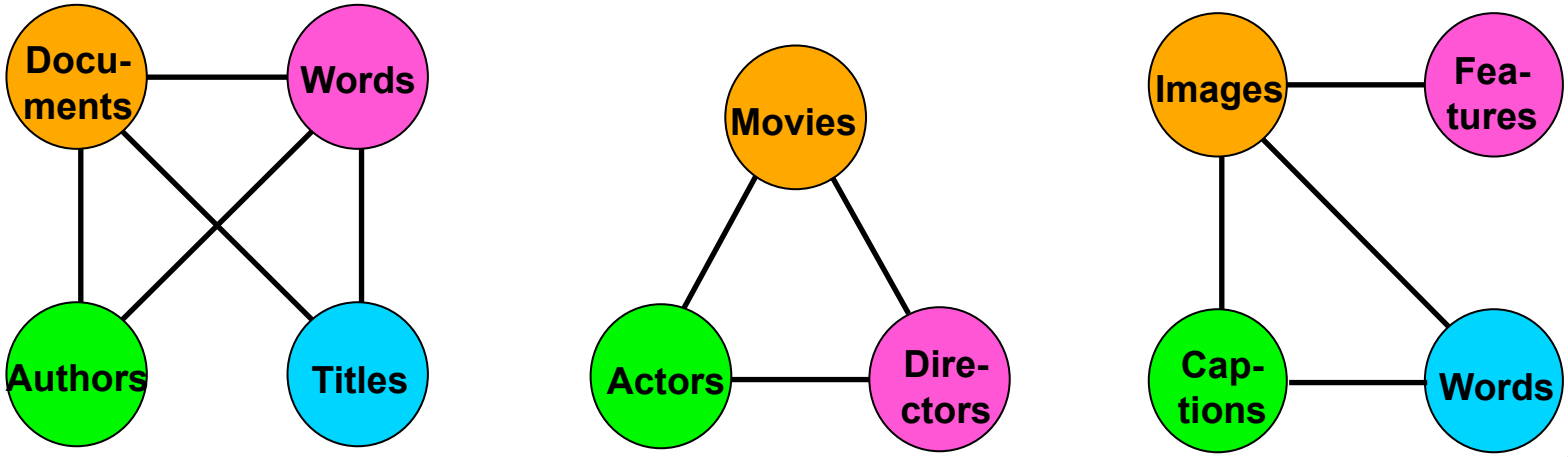
- Simultaneously constructing N clusterings of N modalities of the data
  - Clusterings “bootstrap” each other
- Hot topic in machine learning
  - Dhillon et al. SIGKDD-2003
  - Bickel and Scheffer ICDM-2004
  - Bekkerman et al. ICML-2005
  - And many others

Multi-way distributional clustering

***a.k.a. MDC** (Bekkerman et al. ICML-2005)*

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

Pairwise Interaction Graph

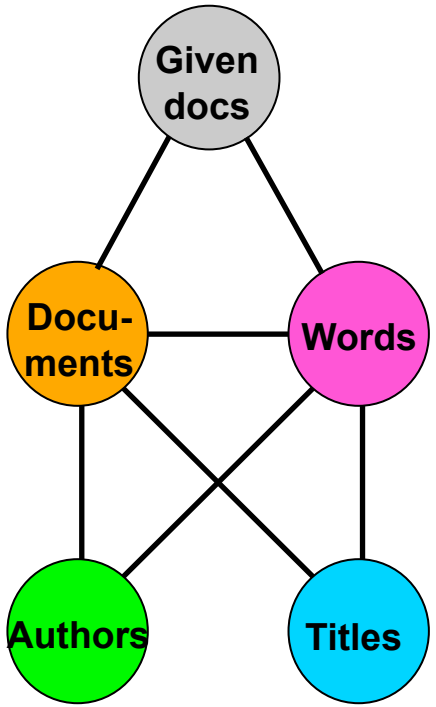


Objective function of MDC

- Let (V, E) be pairwise interaction graph
- **Objective:** sum of pairwise MI
  - $$\max_{\tilde{X}_1,\ldots,\tilde{X}_N} \sum_{(V_i,V_j)\in E} \mathrm{I}(\tilde{X}_i;\tilde{X}_j)$$
  - Subject to $\left|\tilde{X}_i\right| = K_i, \;\; i = 1,\ldots,N$
- No multi-dimensional probability tables
- Can be easily factorized

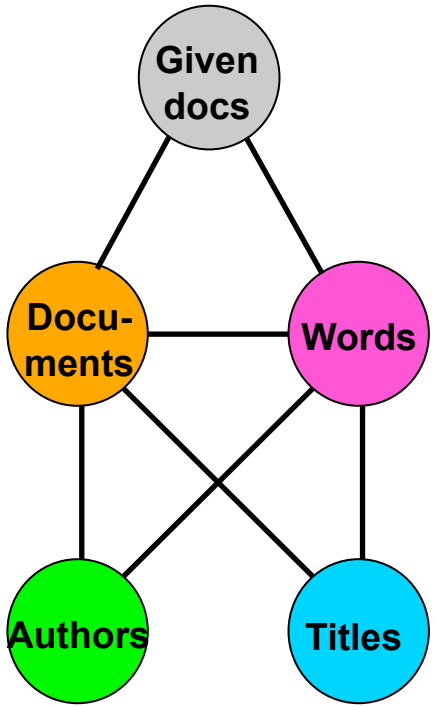
Semi-supervised case

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



Possible solution

- Make “Documents” be a random variable
  - Over all possible partitionings of documents
- “Given docs” will be an *observed* random variable
  - Whose value is a given partitioning



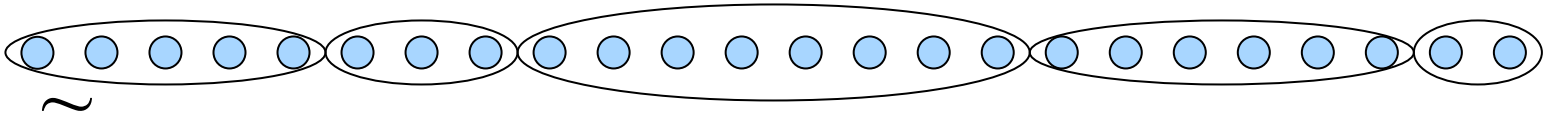
Combinatorial random variable

- 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

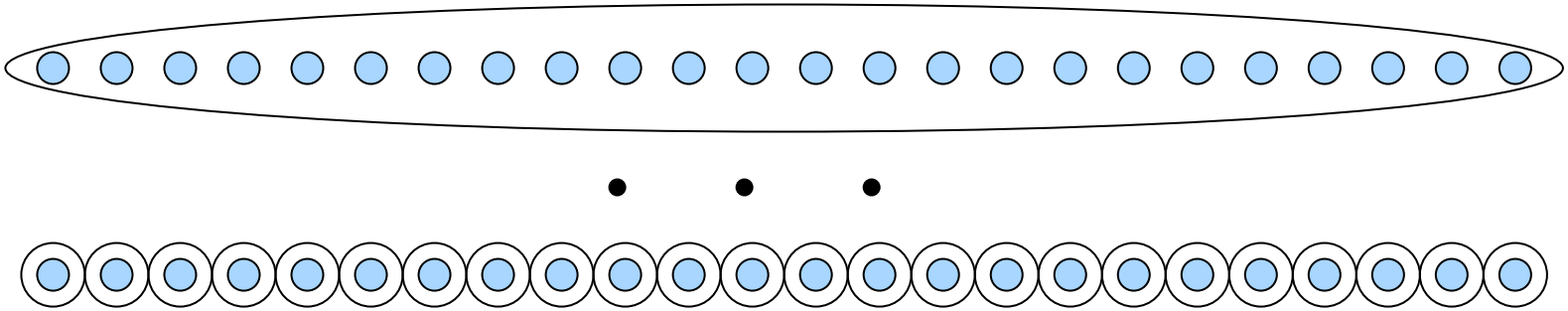
Example: hard clustering

- 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)

- Markov Random Field with combinatorial random variables
- Goal:
  - Find “best” (most likely) assignment to combinatorial random variables
    - *i.e. Most Probable Explanation (MPE)*
- Challenges:
  - Usually, $P(\tilde{X}^c)$ cannot be explicitly specified
  - No existing inference methods applicable

Properties of Comraf models

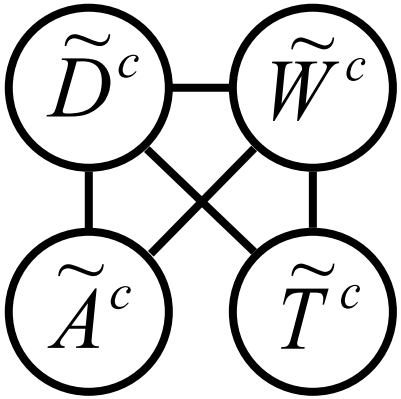
- **Neither generative nor discriminative**
  - No generative assumptions to make
  - No training data required
- **Compact:** one node per “concept”
  - Such as “*clusterings of documents*”, “*rankings of movies*”, “*subsets of images*” etc.
  - Model learning is feasible
- **Generic:** applicable to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

Comraf model

- Graph G over combinatorial r.v.'s
- Objective function F as in MDC
- Important special cases:
  - A "hard" variation of Information Bottleneck (Tishby et al., 1999)
  - Information-theoretic co-clustering (Dhillon et al., 2003)
  - MDC (Bekkerman et al., 2005)

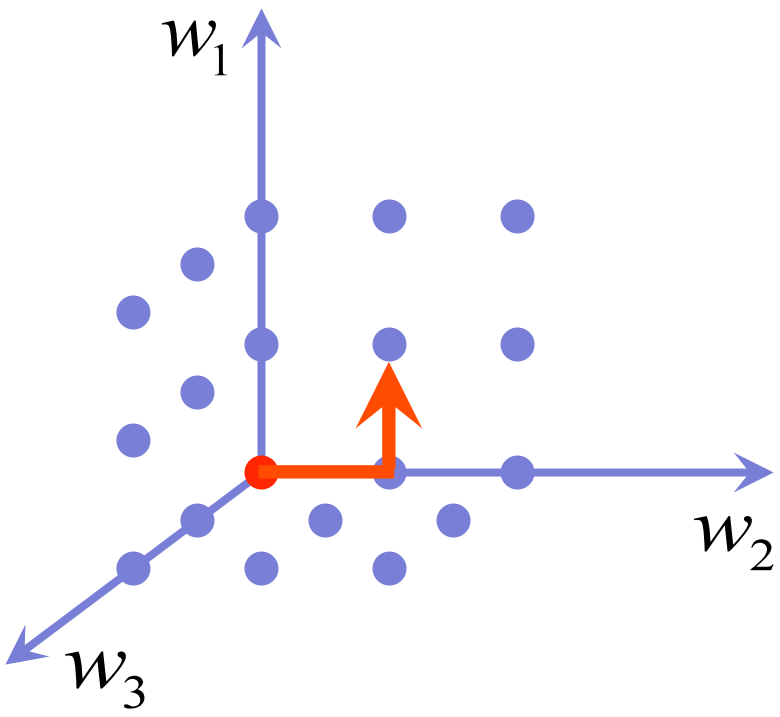
Inference in Comraf models

- *Iterative Conditional Mode (ICM)*
  - Fix current values of all variables but one
  - Optimize this variable wrt its neighbors
  - Fix its new value and move to another variable
  - Round-robin over the variables



Inference: local optimization

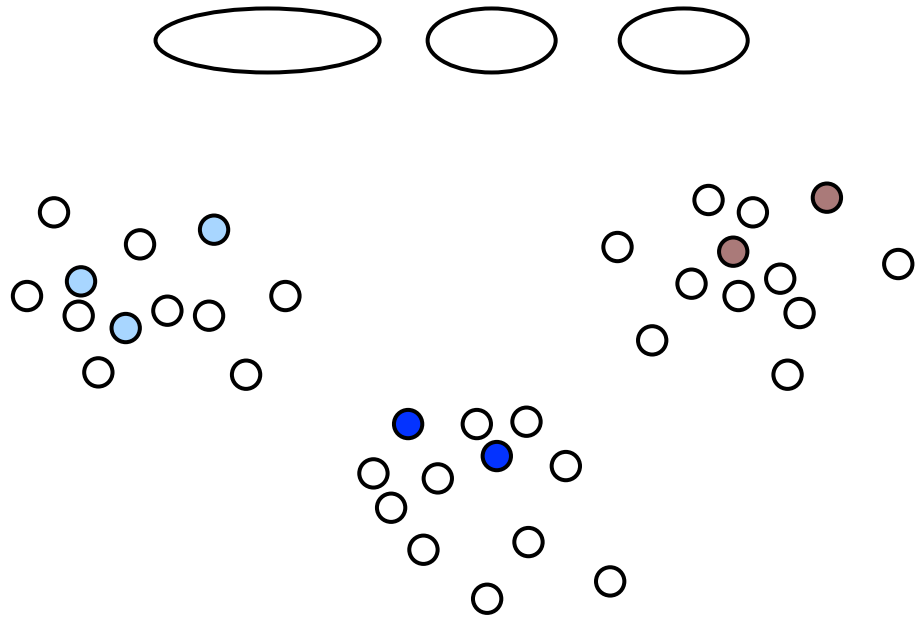
Lattice of possible solutions



- For each variable
- Start with some solution
  - Say, (0,0,0)
    - All data points are in cluster c_0
- Traverse the lattice
  - While maximizing the objective

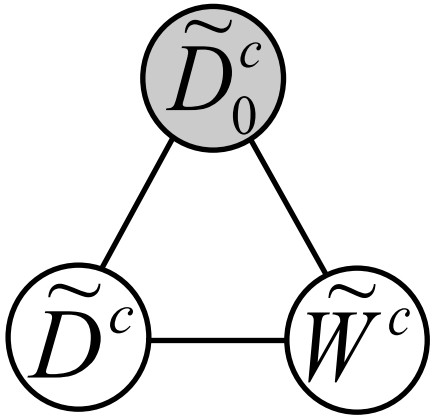
Semi-supervised clustering

- Labeled data compose a natural partitioning



Intrinsic Comraf model

- We are given some labeled documents
  - Which form partitioning $\tilde{d}_0^c$
  - Represented as observed r.v. $\tilde{D}_0^c$
  - With an r.v. $\tilde{D}_0$ defined over $\tilde{d}_0^c$



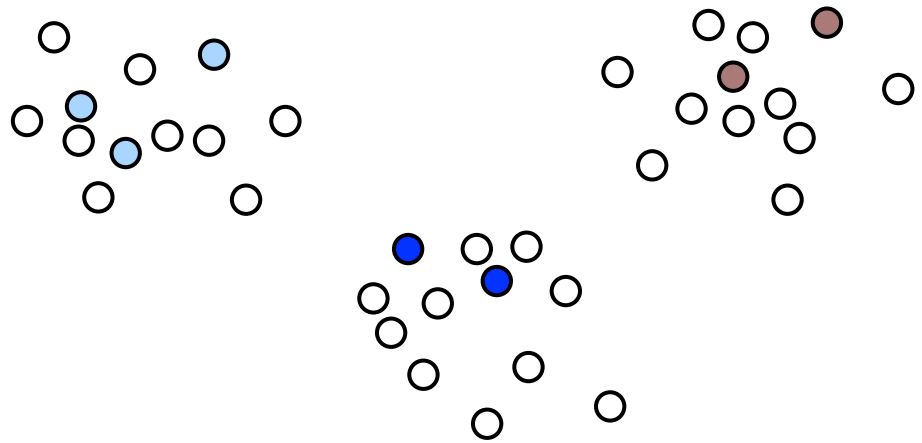
- Objective:

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

- Inference method is the same

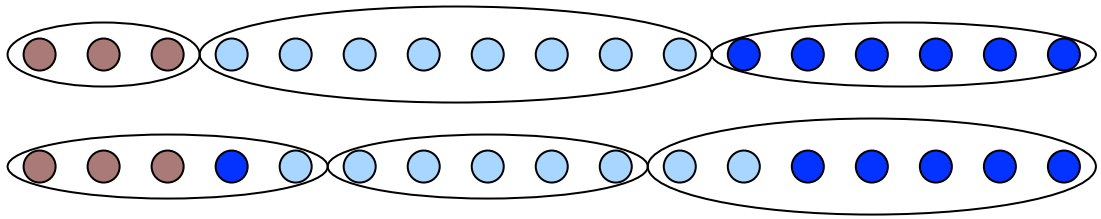
Constrained optimization scheme

- 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:



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

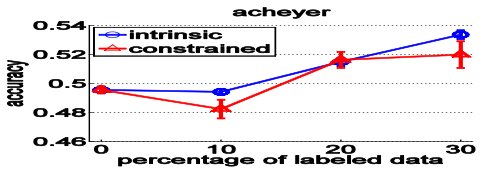
Size of dominant class in cluster c

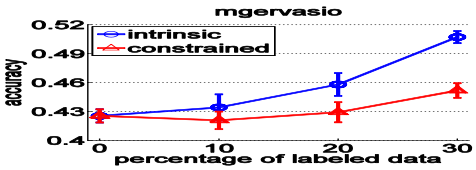
Datasets

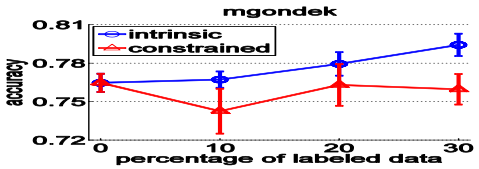
- Three CALO email datasets:
  - acheyer: 664 messages, 38 folders
  - mgervasio: 777 messages, 15 folders
  - mgondek: 297 messages, 14 folders
- Two Enron email datasets:
  - kitchen-l: 4015 messages, 47 folders
  - sanders-r: 1188 messages, 30 folders
- The 20 Newsgroups: 19,997 messages

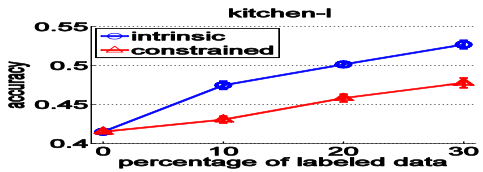
Results on email datasets

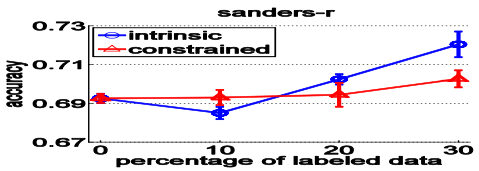
- Randomly choose 10, 20 and 30% of data to be labeled
- Plot the accuracy of the unlabeled portion









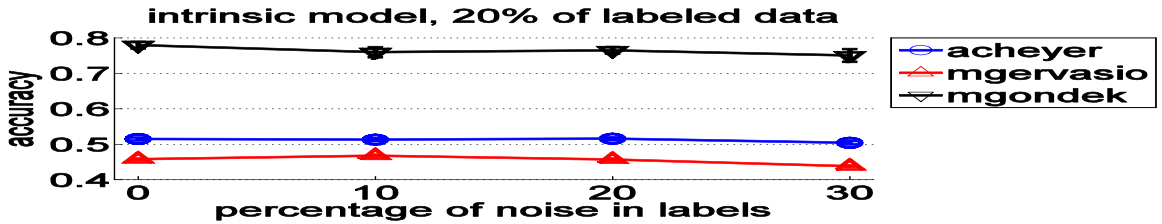


Semi-supervised clustering on 20NG

- 69.5±0.7% unsupervised clustering
  - 57.5% the best previously reported result
- We consider 10% of data as labeled
- 74.8±0.6% constrained scheme
- 78.9±0.8% intrinsic Comraf scheme

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
  - Other applications will also be considered
- The model is generic
  - Semi-supervised case is straightforward
- Inference algorithms are effective
  - And efficient (quadratic)
- Model learning is possible

Thank you!

- The Comraf clustering tool is available at:

http://www.cs.umass.edu/~ronb/mdc.html