# Multi-modal Clustering for Multimedia Collections

Ron Bekkerman,
Jiwoon Jeon

*February 23, 2007*

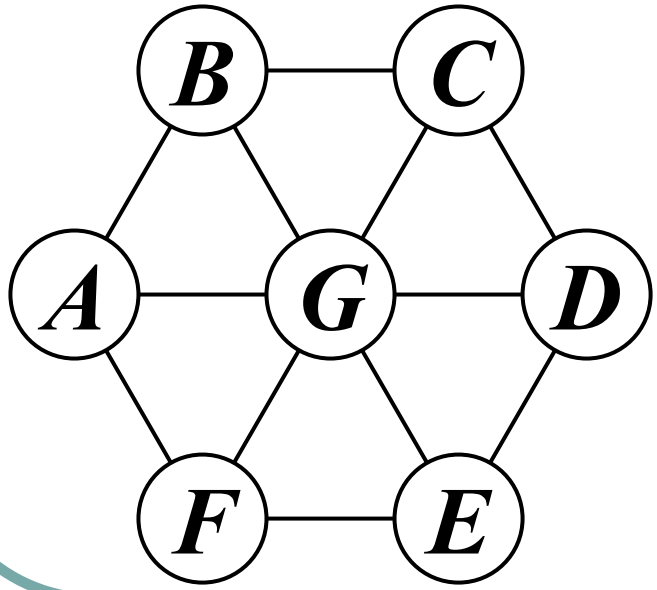
# Motivation

- Multimedia collections are multi-modal
  - Text, images, audio, video are multiple views of the presented concept
- Last year we proposed **Comrafs**
  - A useful model for clustering multi-modal data
- It's a shame not to apply to

  - We focus on clustering images with captions

# Comraf essentials

- Comrafs are Markov Random Fields with nodes of “rich structure”
  - I.e. random variables with very large support
    - Such as “all possible clusterings of a set”

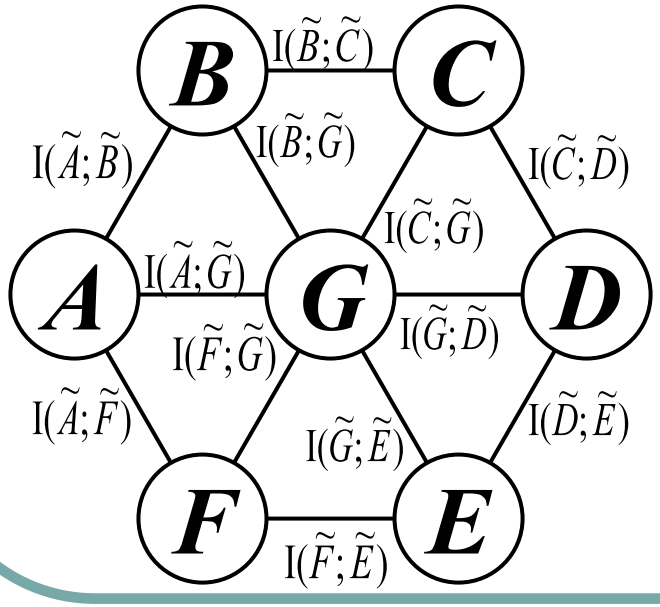



- The goal is to find the best value of each variable
  - Such as “the best clustering”

# Comrafs: objective function

- Best clusterings maximize the objective

$$\mathrm{I}(\tilde{A};\tilde{B})+\mathrm{I}(\tilde{B};\tilde{C})+\mathrm{I}(\tilde{B};\tilde{G})+\mathrm{I}(\tilde{A};\tilde{F})+\mathrm{I}(\tilde{A};\tilde{G})+\mathrm{I}(\tilde{F};\tilde{G})+\mathrm{I}(\tilde{C};\tilde{G})+\mathrm{I}(\tilde{G};\tilde{E})+\mathrm{I}(\tilde{F};\tilde{E})+\mathrm{I}(\tilde{G};\tilde{D})+\mathrm{I}(\tilde{C};\tilde{D})+\mathrm{I}(\tilde{D};\tilde{E})$$

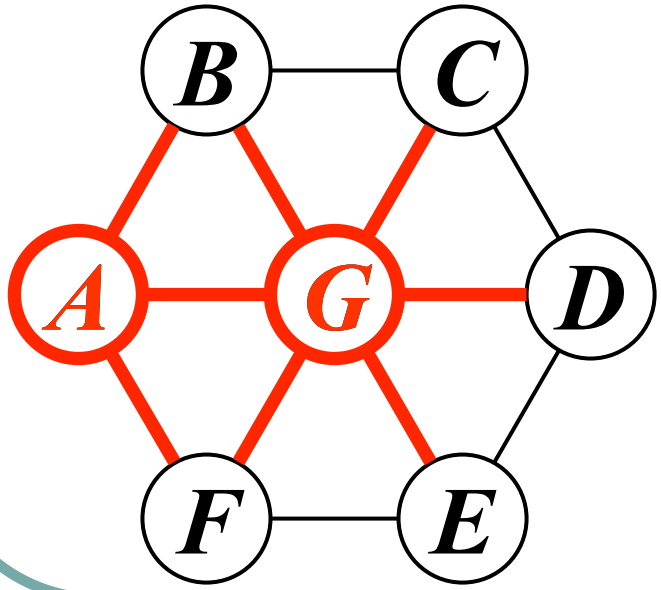



- A potential is defined on each edge

# Comrafs: inference procedure

- Best clusterings maximize the objective

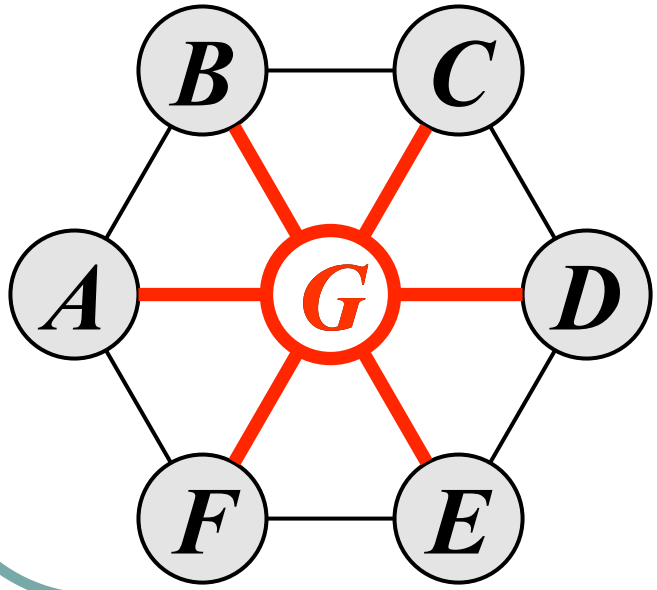
$$\mathrm{I}(\tilde{A};\tilde{B})+\mathrm{I}(\tilde{B};\tilde{C})+\mathrm{I}(\tilde{B};\tilde{G})+\mathrm{I}(\tilde{A};\tilde{F})+\mathrm{I}(\tilde{A};\tilde{G})+\mathrm{I}(\tilde{F};\tilde{G})+\mathrm{I}(\tilde{C};\tilde{G})+\mathrm{I}(\tilde{G};\tilde{E})+\mathrm{I}(\tilde{F};\tilde{E})+\mathrm{I}(\tilde{G};\tilde{D})+\mathrm{I}(\tilde{C};\tilde{D})+\mathrm{I}(\tilde{D};\tilde{E})$$




- Fix values of all nodes but one
- Optimize the node wrt its Markov blanket
- Move to another node

# Clustering in multimedia

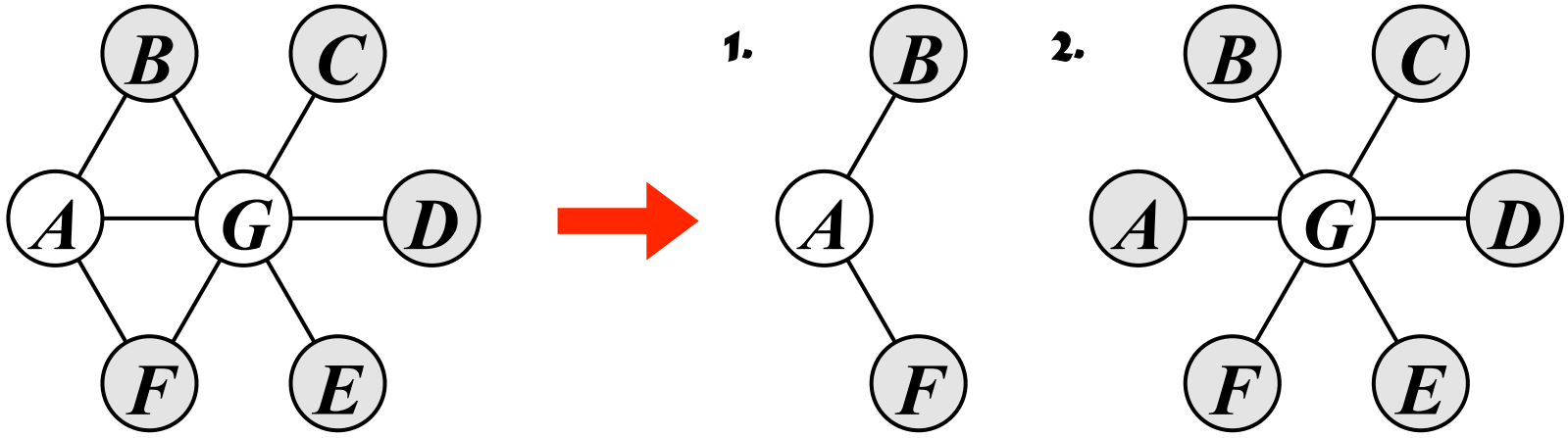
- Many views are dense enough
  - Such as *colors*: no need to cluster them
    - Even *caption words* may not be clustered



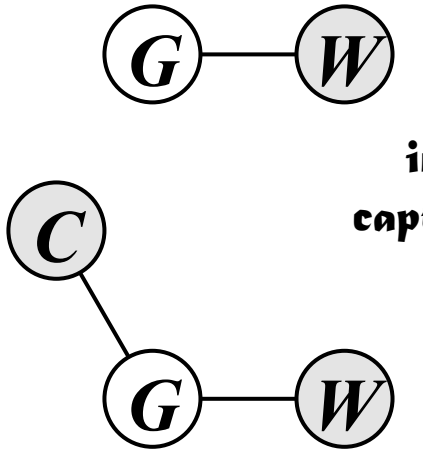

- We end up with one *target* node ***G***
  - And *observed* nodes
- Observed nodes do not interact with each other

# Comraf* models

- Comraf models of an asterisk topology
  - With observed nodes around the target node
- A general Comraf can be translated into a sequence of Comraf*



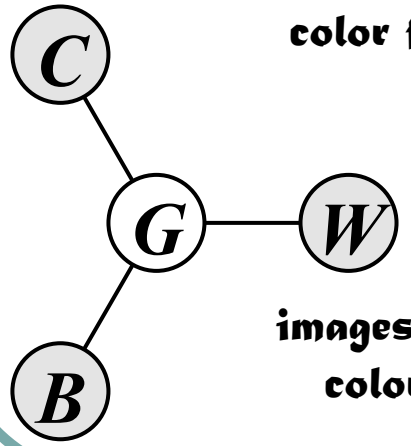
# Particular models



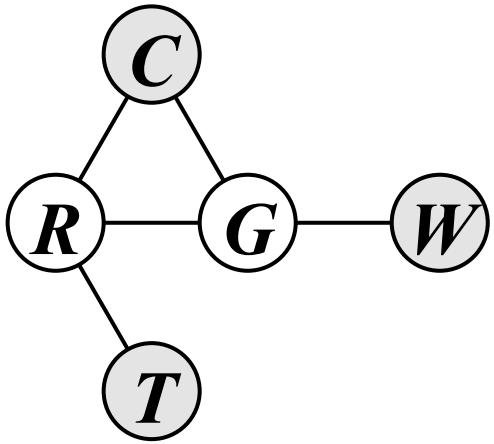

images / caption words



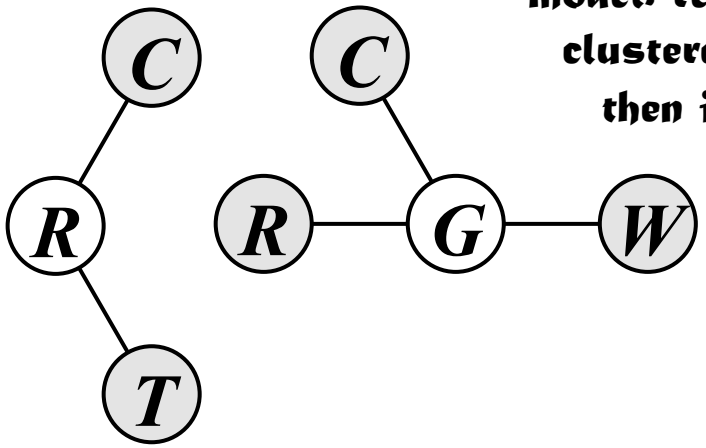

images / words / color frequencies




images / words / colors / blobs




A general Comraf model: images / words / colors / regions / texture



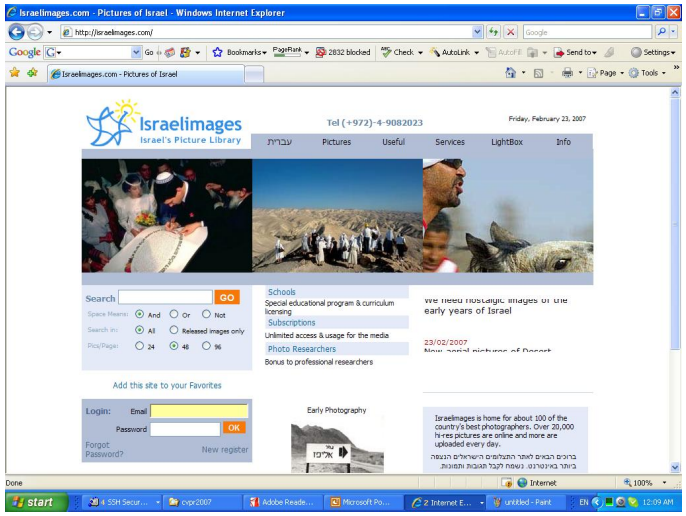

2-step Comraf* model: regions are clustered first, then images

# Image processing glossary

- *Blobs*: clusters of image regions
  - Roughly correspond to words in text
  - We use an existing set of blobs
- *Regions*: rectangular segments of images
  - We use 24 regions
- *Texture*: Gabor features
  - Directions and scales of major activity
  - We use 12 Gabor features
    - 4 directions and 3 scales

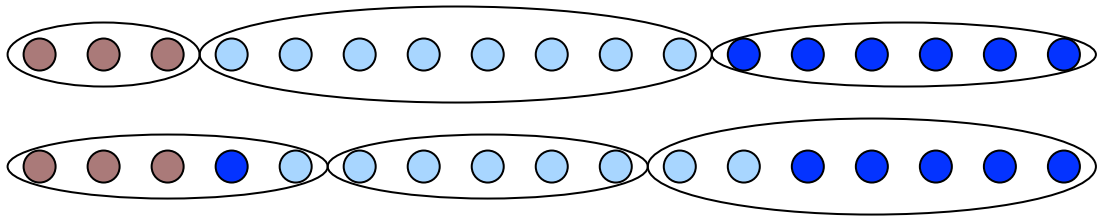
# Datasets

- Corel
  - A benchmark dataset for image processing
  - A subset of 4500 images, 50 categories
- Israel Images
  - Collected especially for this project ☺
  - 1823 images, 11 categories



# Evaluation methodology

- Clustering evaluation
  - Is generally unintuitive
  - Is an entire research field
- We use the *clustering accuracy* measure
  - One of the standard measures available
  - Ground truth:
  - Our results:

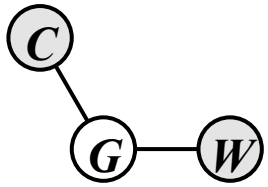


$$Acc = \frac{1}{|G|}\sum_{c}\gamma_c$$

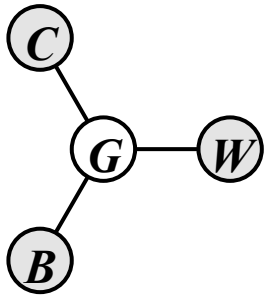
Size of dominant class in cluster $c$

# Results on Israel Images

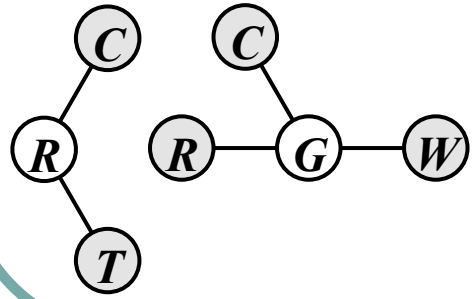

- 44.2 ± 1.0%




- 54.2 ± 0.9%




- No blob data available

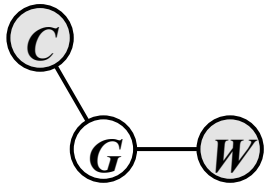



- 68.8 ± 0.9%

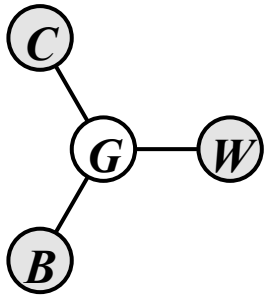
k-means: 22%

# Results on Corel

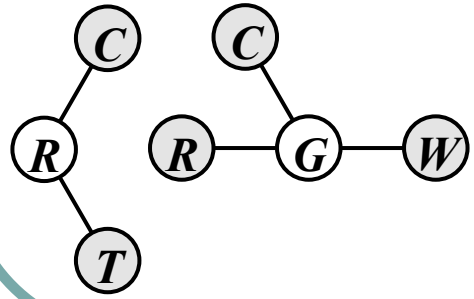

- 46.6 ± 0.5%




- 55.3 ± 0.5%



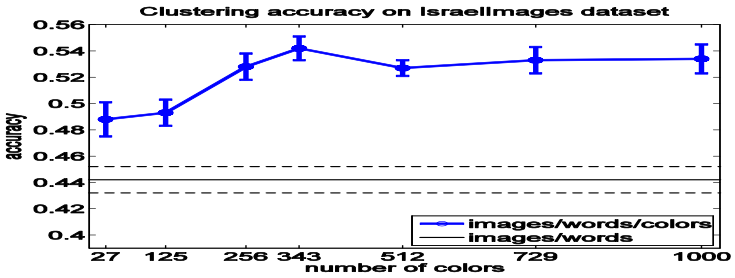

- 60.1 ± 0.3%

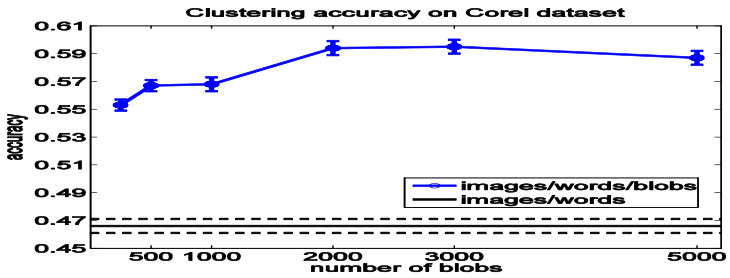



- 61.2 ± 0.4%

k-means: 22%

# Good number of colors / blobs



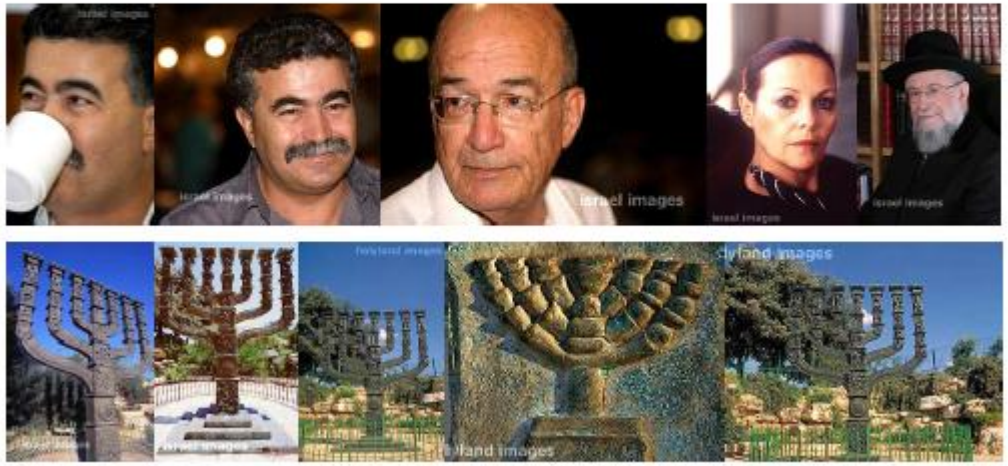
# A Corel example

(a) Clustering results using only caption words, Corel dataset

(b) Clustering results using words and blobs, Corel dataset

# An Israel Images example

(a) Clustering results using only caption words, IsraelImages dataset



(b) Clustering results using words and color histograms, IsraelImages dataset

# Conclusion

- Comrafs are very useful for clustering multimedia
- A lot of experiments still to conduct
- A lot of design choices still to make