Probabilistic Models in Social Network Analysis

Sargur N. Srihari

University at Buffalo, The State University of New York

USA

US-India Workshop on
Large Scale Data Analytics and Intelligent Services
December 2011

Plan of Presentation

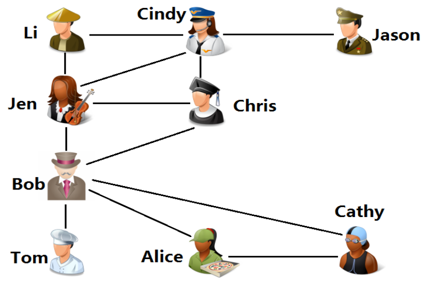
- SNA and PGMs
- Science
- Applications
- Research on Learning Models
- Scientific Challenges
- Application Challenges

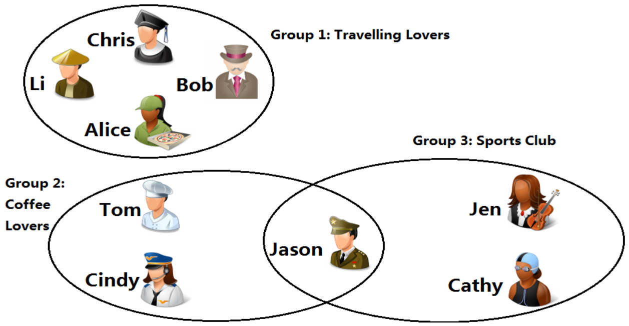
SNA and PGMs

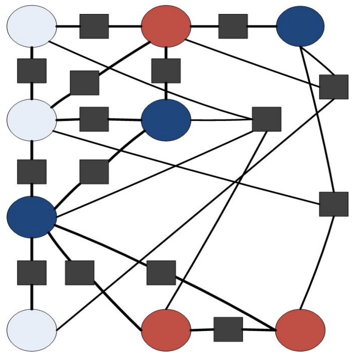
1. SNA: Explosive growth in SN digital data
    1. OSNs of kinship, email, and affiliation groups
    2. Mobile communication devices, bibliographic citations, business interactions
2. PGMs: compact aggregate representations
    - Essential for many variables
        - Otherwise data requirements impossible
        - Inference intractability is still an issue but approximated
    - Automatic learning necessary since structure and data continually change
        - Area of ML little explored

Application: Latent Variable Prediction

- Social network Data Graph
  - Actors, pairwise links
- Affiliation network Data Graph
  - Societies, complete links
- Combined: social-affiliation graph
- Markov Network Variable Graph
  - Bipartite graph
  - Predict nodes Y given nodes X



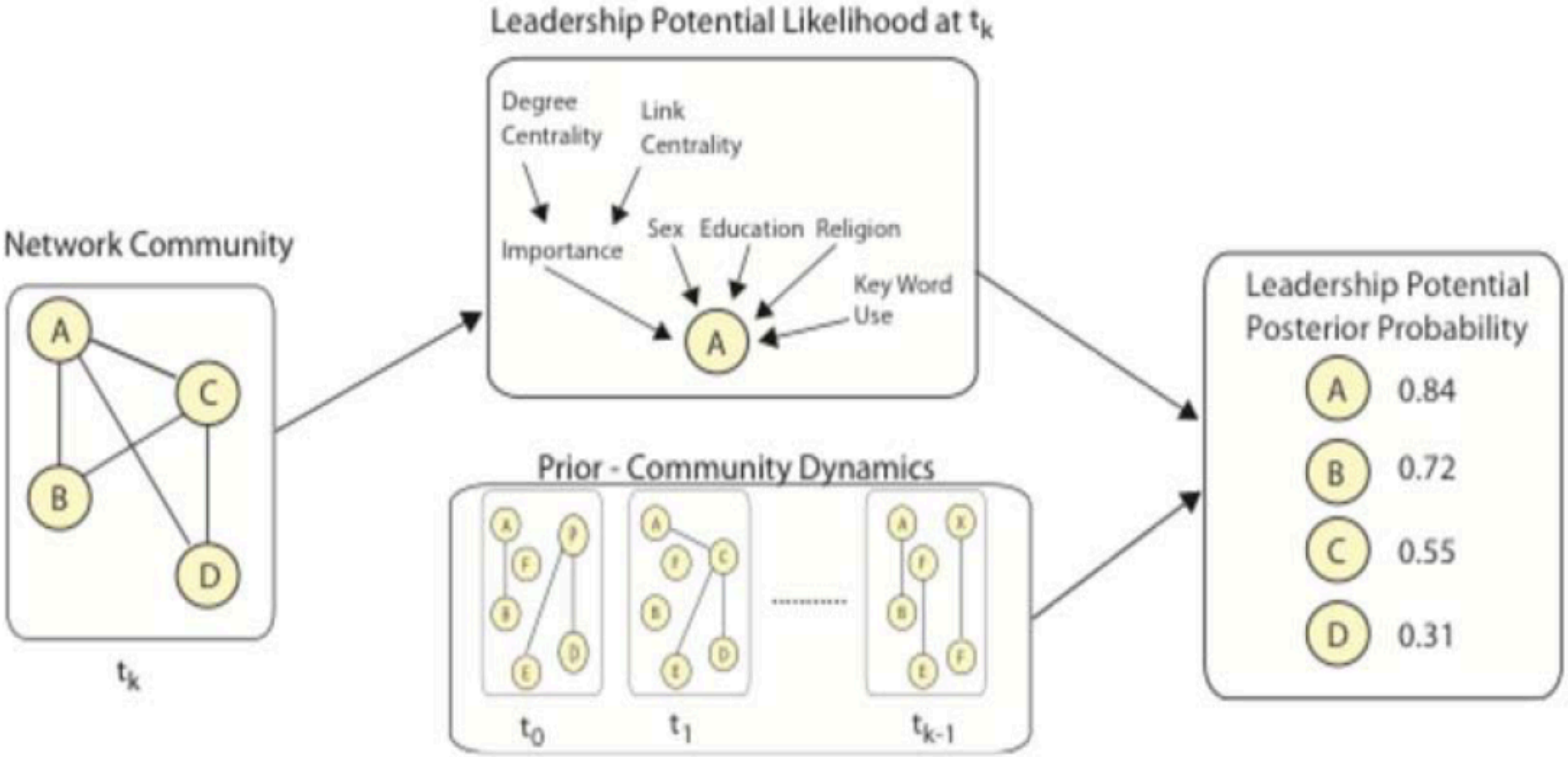




Application: Leadership potential

(Joint work with Rachael Blair)

Importance determined by centrality measures and attributes

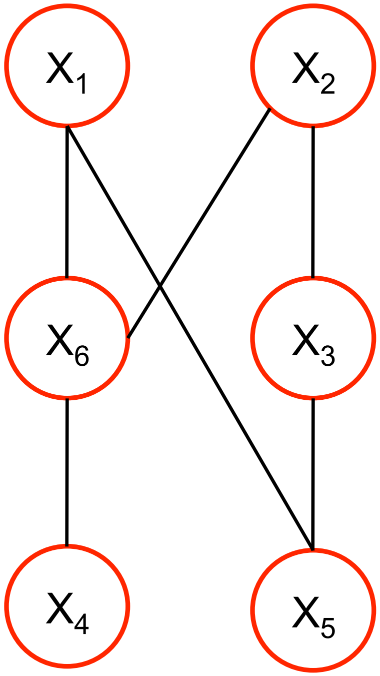


Research: Learning PGMs Taxonomy

1. Data Sampling
2. Parameter Learning (given structure)
    1. BN: Straight-forward since local CPDs
    2. MN: Global coupling and no closed-form solutions
3. Structure Learning
    1. Search through network space
4. Partial Data
    1. EM

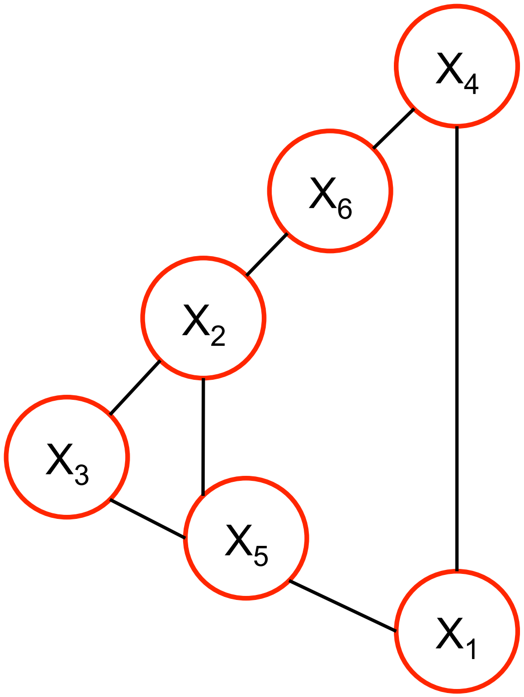
Structure learning (Baseline Models)

(Joint work with Dmitry Kovalenko)



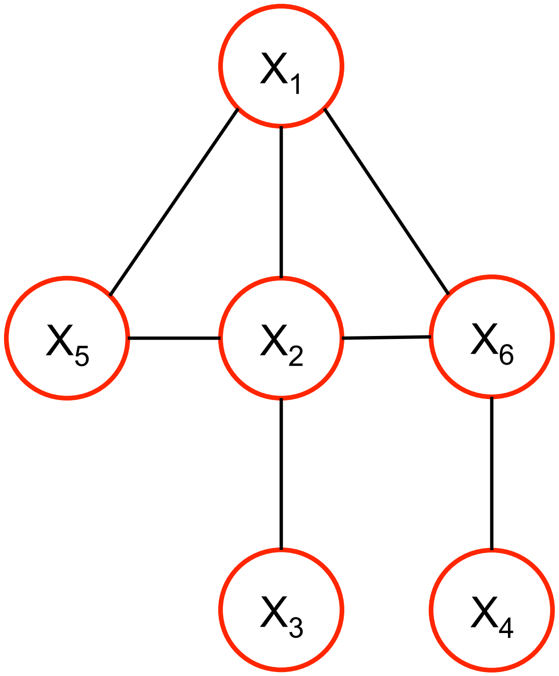
MN_1

- Designed “by hands“



MN_2

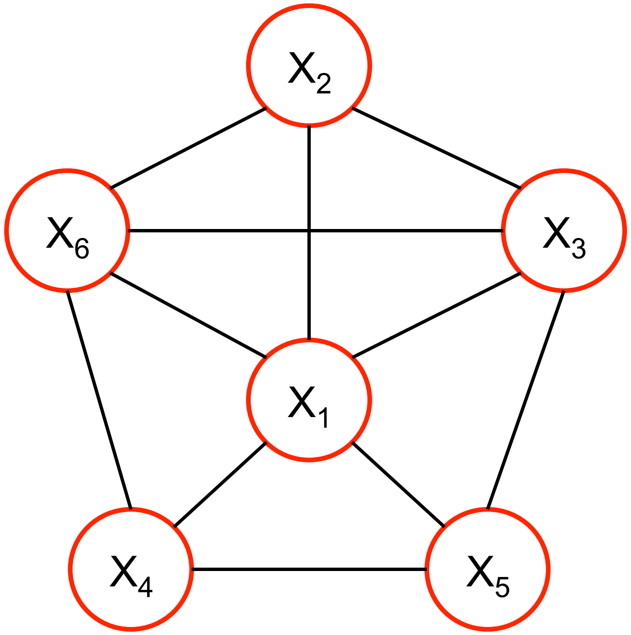
- Automatically designed as a BN



MN_3

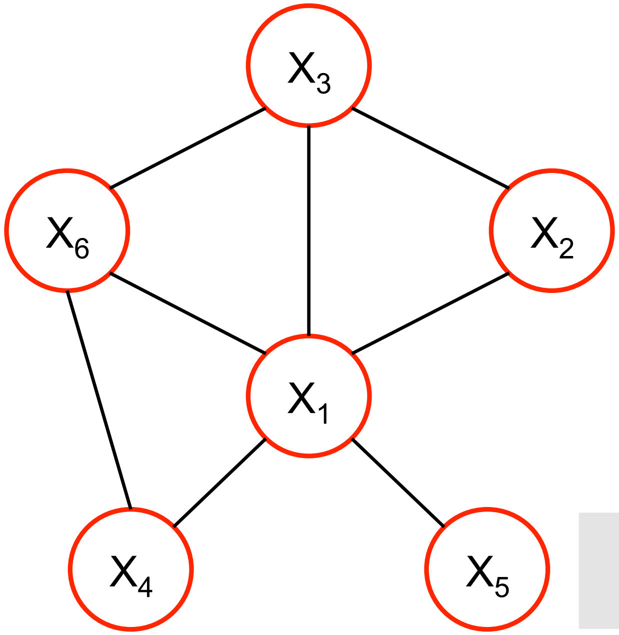
Modified Chow-Liu algorithm (NIPS 2011)

Performance of Competitive Models



MN_4

Designed by L-BFGS optimization
with L_1-regularized
log likelihood
#parameters=156



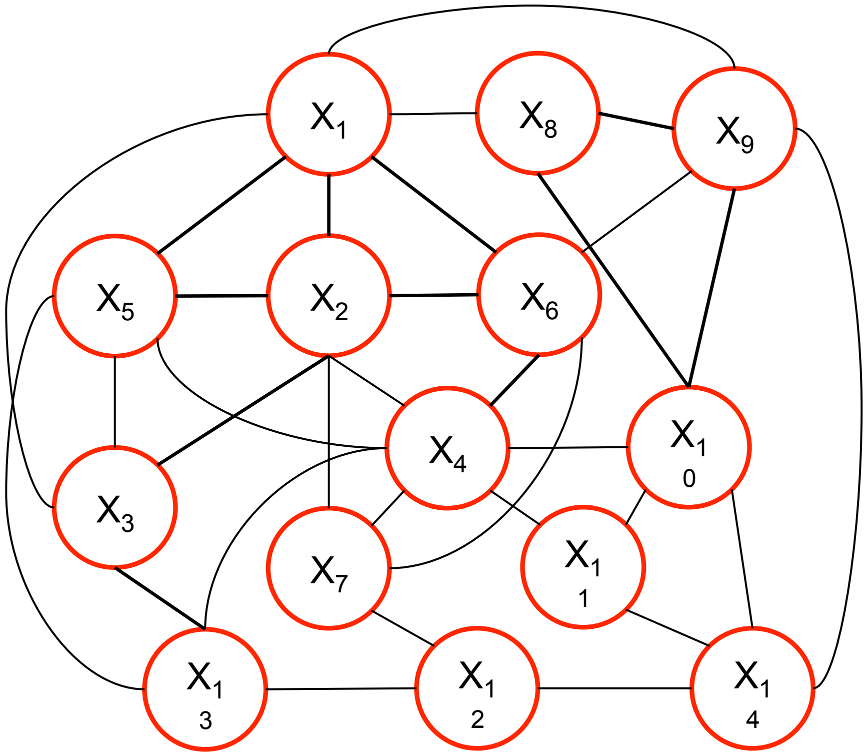
MN_5

Designed by CASL
# of parameters: 126

Cross-validation
Log-loss

Markov Network	Average log-loss
MN_1	2687
MN_2	2702
MN_3	2686
MN_4	2645
New MN_5	2649

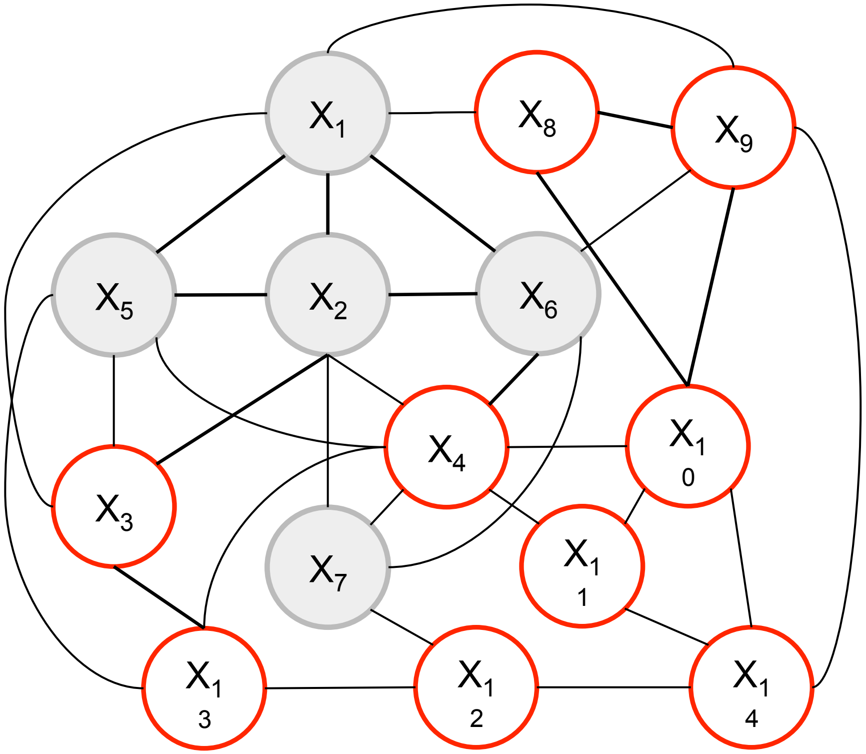
Suggested way of running algorithm on social network data sets



Social Network

1. Pick a number:
   N – size of a feature
2. One by one get different connected subgraphs of size N (create a set M)
3. Run algorithm on data set M
4. Create a new set T
5. Test MRF on T

Suggested way of running algorithm on social network data sets



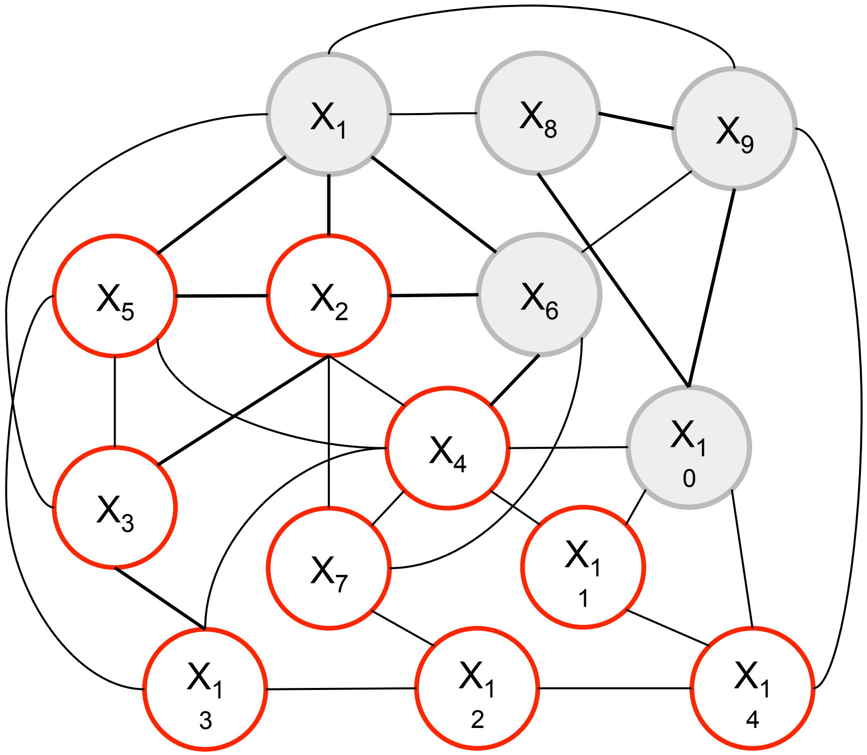
Social Network

1. Pick a number:

   N – size of a feature

2. One by one get different connected subgraphs of size N (create a set M)
3. Run algorithm on data set M
4. Create a new set T
5. Test MRF on T

Suggested way of running algorithm on social network data sets



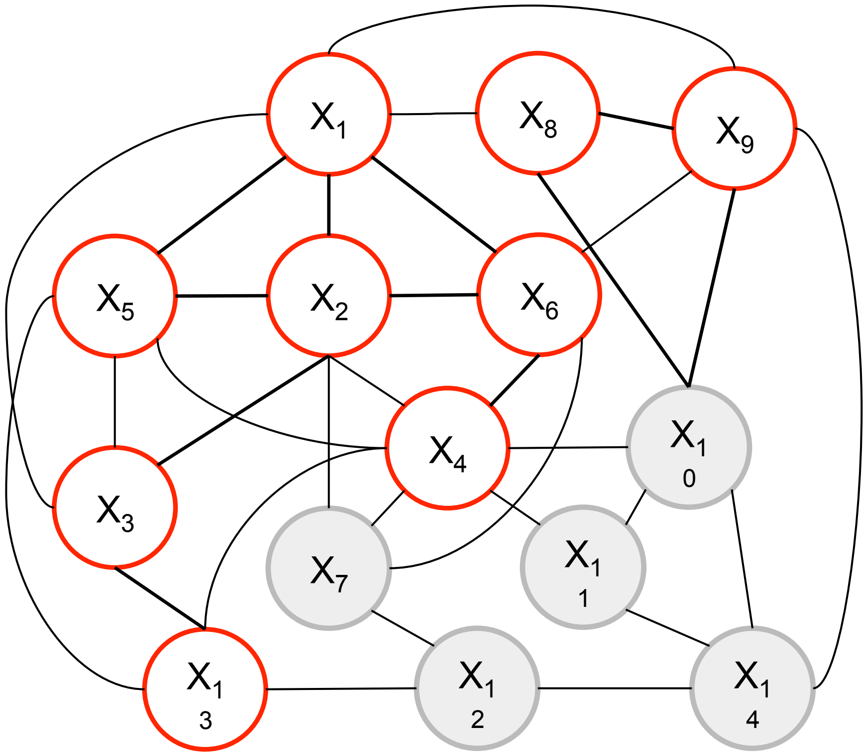
Social Network

1. Pick a number:

   N – size of a feature

2. One by one get different connected subgraphs of size N (create a set M)
3. Run algorithm on data set M
4. Create a new set T
5. Test MRF on T

Suggested way of running algorithm on social network data sets



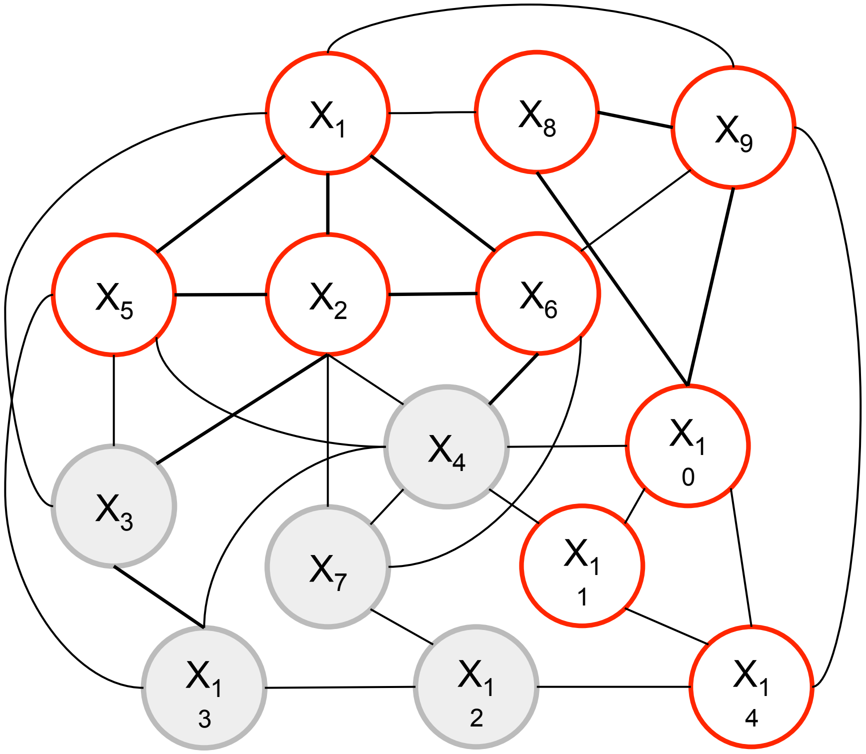
Social Network

1. Pick a number:

   N – size of a feature

2. One by one get different connected subgraphs of size N (create a set M)
3. Run algorithm on data set M
4. Create a new set T
5. Test MRF on T

Suggested way of running algorithm on social network data sets



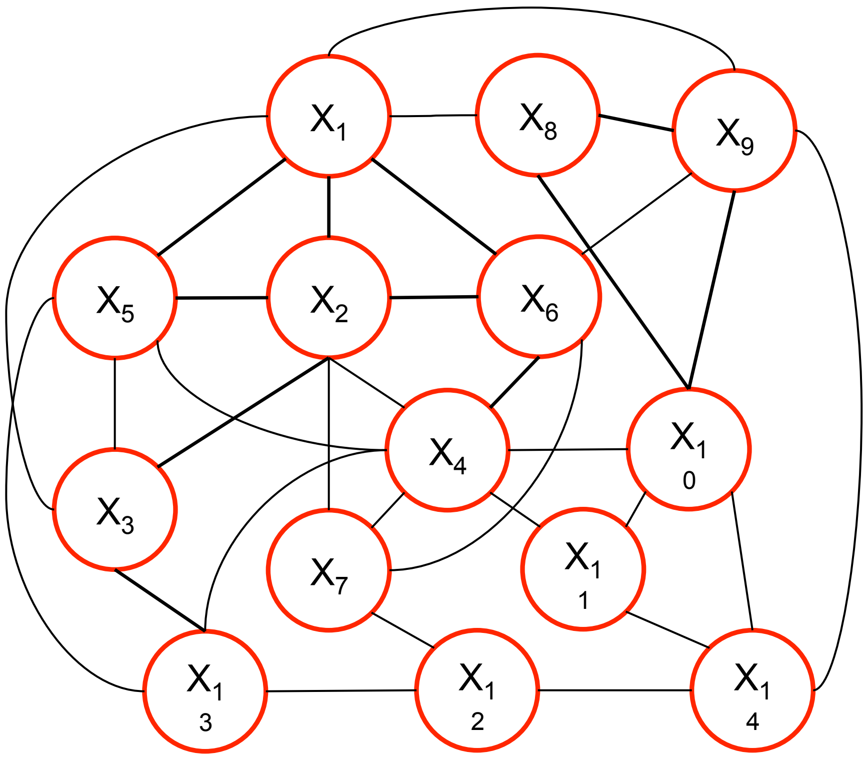
Social Network

1. Pick a number:

   N – size of a feature

2. One by one get different connected subgraphs of size N (create a set M)
3. Run algorithm on data set M
4. Create a new set T
5. Test MRF on T

Running Structure algorithm on social network data sets

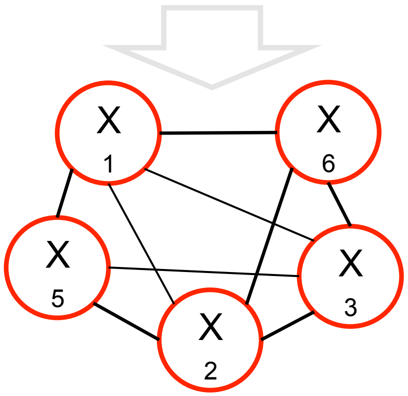


Social Network

1. Pick a number:

   N – size of a feature

2. One by one get different connected subgraphs of size N (create a set M)
3. Run algorithm on data set M



MRF

Data Sets for Research

Facebook datasets were collected in April of 2009:

- MHRW - A sample of 957K unique users obtained Facebook-wide by 28 independent Metropolis-Hastings random walks
- UNI - A sample of 984K unique users that represents the "ground truth" i.e., a truly uniform sample of Facebook userIDs, selected by a rejection sampling from the system's 32-bit ID space.

There are 2 files for each dataset:

- <uid> <#times sampled> <friend_uid_1> <friend_uid_2> .. <friend_uid_j>
- <uid> <#times sampled> <#totalfriends> <privacy settings> <networkID(s)>

There is no attribute info in the data.

The privacy settings consist of four basic binary privacy attributes:
1) Add as friend 2) Photo thumbnail 3) View friends 4) Send message

Mobile Data Challenge (by Nokia)

Released in Week 1, 2012.

Contains data of 200 users for more than 1 year, its features are:

- Phone usage (full call and message log)
- Phone status data (GPS readings, operation mode)
- Environment data (accelerometer samples, wi-fi access points, bluetooth devices)
- Personal data (full contact list, calendar)
- Users info (gender, age, occupation, marital status, occupation etc.)

Next Set of Challenges- Scientific

- Automatic construction of a generative graphical model for social network (interpreting links as variables that take values from {0,1})
- Dynamical MRF construction for temporal modeling of social networks
- Improving of inference and group selection procedures using existing approach for MN structure construction