Community structure analysis in social networks with relational latent feature models

Manfred Jaeger, Jiuchuan Jiang

Machine Intelligence Group
Department of Computer Science
Aalborg University

Introduction

Numerical Inputs in RBNs

Application: Community Structure Analysis

Hybrid Models

A continuous model:

$$Y \approx \alpha + \boldsymbol{X} \cdot \boldsymbol{\beta} + N(0, \sigma^2)$$

Continuous ingredients:

- Y: **response** (random) variable
- $\boldsymbol{X}$: **predictor** variables (random or not)
- $\alpha, \boldsymbol{\beta}, \sigma^2$: **parameters**

Hybrid Models

A continuous model:

$$Y \approx \alpha + \boldsymbol{X} \cdot \boldsymbol{\beta} + N(0, \sigma^2)$$

Continuous ingredients:

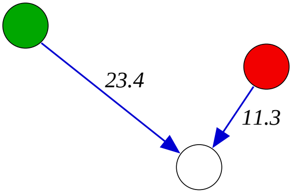
- Y: **response** (random) variable
- $\boldsymbol{X}$: **predictor** variables (random or not)
- $\alpha, \boldsymbol{\beta}, \sigma^2$: **parameters**

Hybrid SRL models:

	Continuous Predictors	Continuous Response	Inference
Hybrid MLN [Wang,Domingos 2008]	Yes	Maybe	approx.
Hybrid ProbLog [Gutmann et al. 2011]	No	Yes	exact
Hybrid Relational Dependency Networks [Ravkic, Ramon, Davis 2015]	Yes	Yes	approx.

Continuous Predictors for SRL

Allow **numerical input relations**:



Examples:

- modelling *discrete sensor states*, **given** the *distances* between the sensors
- modelling of a *friendship* relation, **given** *age* and *income* attribute
- modelling of a *friendship* relation, **given** *latent community membership degrees*

Introduction

Numerical Inputs in RBNs

Application: Community Structure Analysis

RBN Language

Sensor Network Example

```
polluted(S) ← WIF   0.6
              THEN  COMBINE WIF  polluted(X)
                            THEN 0.4
                            ELSE 0.0
                    WITH n-or
                    FORALL X
                    WHERE upstream(X, S)
              ELSE  0.2;
```

Noisy-or and other combination functions map *multisets of probability values* to a probability value.

RBN language:

- Inductively defined from *constants*, *logical atoms*, `WIF-THEN-ELSE`, and `COMBINE-WITH-FORALL-WHERE` constructs.
- Boolean relations interpreted numerically (0,1-valued)
- Uniform treatment of parameters and logical atoms as probability (sub-) formulas.

Adding Numeric Inputs

Syntax

- No change: atomic expressions may now also stand for numeric relations

Adding new combination functions

- Add *logistic regression* combination function

$$\textit{l-reg}\{p_1, \ldots, p_n\} := e^{\sum p_i}/(1 + e^{\sum p_i})$$

- Takes multiset of real numbers and returns probability

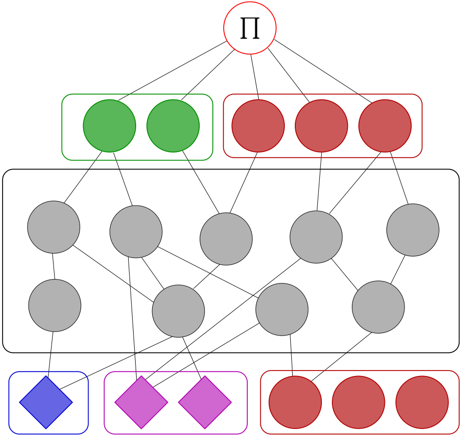
Example

$\textit{polluted}(S) \leftarrow$ WIF 0.6
THEN COMBINE WIF $\textit{polluted}(X)$
THEN $1/\textit{distance}(X, S)$
ELSE 0.0
WITH l-reg
FORALL X
WHERE $\textit{upstream}(X, S)$
ELSE 0.2;

Likelihood Graph

Inference and Learning:

- Numeric input relations treated like parameters
- Learning by gradient ascent using *likelihood graph* data structure



Data: ground atoms
observed, unobserved

Ground sub-formula nodes

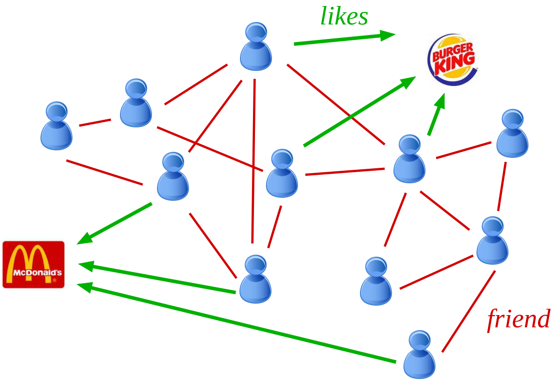
Inputs:
Parameters: Model, Num. Atoms
Unobserved atom indicators

Introduction

Numerical Inputs in RBNs

Application: Community Structure Analysis

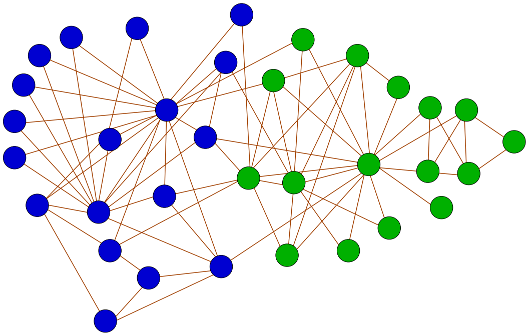
Social Networks



- objects of different types
- connected by different relations

Communities

Classic: community structure discovery as graph partitioning:



Zachary karate club network partitioned into two communities [Newman, Girvan 2004]

- Graded community membership degrees by *soft clustering*
- Little work on community detection in multi-relational networks

Latent Feature Model

Idea:
Obtain a more fine-grained picture of community structure by *community centrality degrees* , that

- reflect how well connected a person is with each community
- are not normalized to sum to one

Latent Feature Model

Idea:
Obtain a more fine-grained picture of community structure by *community centrality degrees* , that

- reflect how well connected a person is with each community
- are not normalized to sum to one

Relational probabilistic formalization:

- Introduce a set $\{C_1, \ldots, C_N\}$ of community objects
- Introduce latent binary numeric relation between nodes V and communities C:

$$u(V, C)$$

Interpretation: $u(V, C)$ is community centrality degree of V within C.

- With each relation r_j associate a latent attribute of community objects

$$t_j(C)$$

Interpretation: $t_i(C) \sim$ affinity of objects in community C to form links of type r_i

Latent Feature Model

Idea:
Obtain a more fine-grained picture of community structure by *community centrality degrees* , that

- reflect how well connected a person is with each community
- are not normalized to sum to one

Relational probabilistic formalization:

- Introduce a set $\{C_1, \ldots, C_N\}$ of community objects
- Introduce latent binary numeric relation between nodes V and communities C:

$$u(V, C)$$

Interpretation: $u(V, C)$ is community centrality degree of V within C.

- With each relation r_j associate a latent attribute of community objects

$$t_j(C)$$

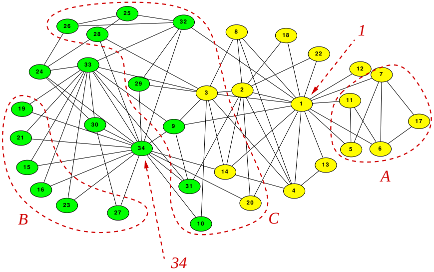
Interpretation: $t_i(C)$ $\sim$ affinity of objects in community C to form links of type r_i

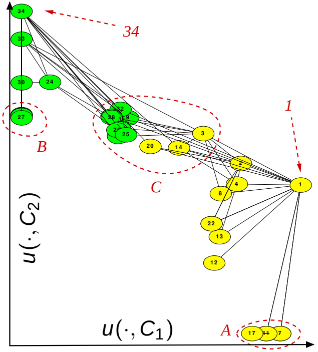
Latent Feature Model

$$P(r_i(V, W)) \leftarrow l - reg(\alpha_i + \sum_{C:community(C)} u(V, C) \cdot u(W, C) \cdot t_i(C))$$

Zachary Result

Learned u-values for Zachary ($N = 2$; one relation; no t_i-parameters):





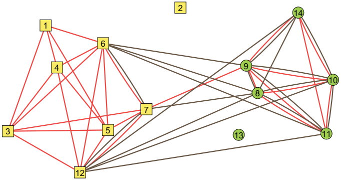
- Maximal $u(\cdot, C_i)$ values identify community centers of C_i
- Important *bridge* nodes between communities characterized by large sum of $u(\cdot, C_i)$ values.
- Learn identical $u(\cdot, C_i)$ values for structurally indistinguishable nodes

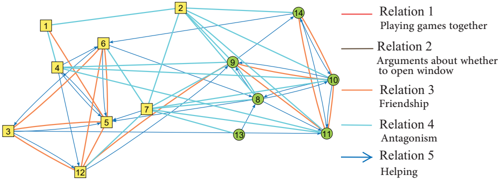
5679 nodes in LG; time per restart: 31s

Multi-Relational Networks

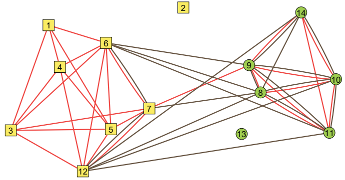
The wiring room network:

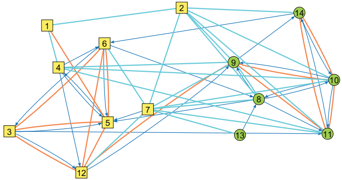
- 14 persons
- 5 relations
  - 3 positive, 1 antagonistic, 1 ambivalent
  - 4 undirected, 1 directed

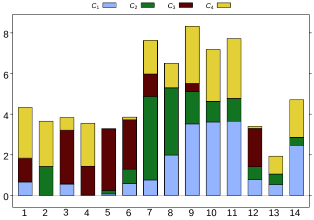




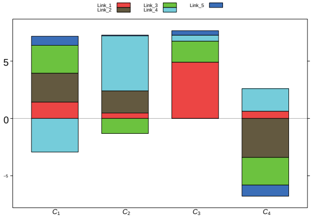
Wiring Room Results







Learned $u(V, C)$ values



Learned $t_i(C)$ values

Community significance values: C_3 : 71.4, C_1 : 62.0, C_2 : 39.5, C_4 : 14.4.

Managing Complexity

Size of likelihood graph is quadratic in number of network nodes.

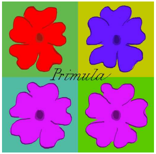
- Current limit: about 100 nodes

Strategies:

- Iteratively construct partial likelihood graphs only
  - for a subset of the parameters and numerical atoms (block gradient ascent)
  - for a subset of the data (stochastic gradient ascent)
- Sub-sample the *false* links

Conclusion

- Integrating continuous variables into probabilistic relational models can be hard, but adding numeric **input (predictor)** relations into RBNs comes almost for free
- Supports construction of standard, interpretable (causal) models
- All implemented in public *Primula* toolbox (and available in next release ...):



- Applied to community structure analysis in (social) networks:
  - new model with latent feature variables representing *community centrality degrees*
  - discovery and characterization of communities in multi-relational networks
- General purpose SRL toolbox instrumental for experimentation with alternative models
- For "industrial strength" use of final model: custom-built learner may be needed