CS6604 Digital Libraries

Social Communities Knowledge Management: Social Interactome

Final Term Project Presentation

Presenter

Prashant Chandrasekar

{peecee}@vt.edu

Instructor

Dr. Edward A. Fox

Virginia Polytechnic Institute and State University

Blacksburg, VA, 24061

May 2, 2017

Acknowledgements

- Dr. Edward A. Fox
- Global events team
- Social Interactome team
  - The Social Interactome of Recovery: Social Media as Therapy Development (NIH Grant 1R01DA039456-01)
- Xuan Zhang and Yufeng Ma
- Mostafa Mohammed

Outline

- Background
  - Social network community; Social Interactome
  - Data
- Challenges
- Goal
- Approaches
- Network Classification
- Learning via Markov Logic Networks
- Future Work

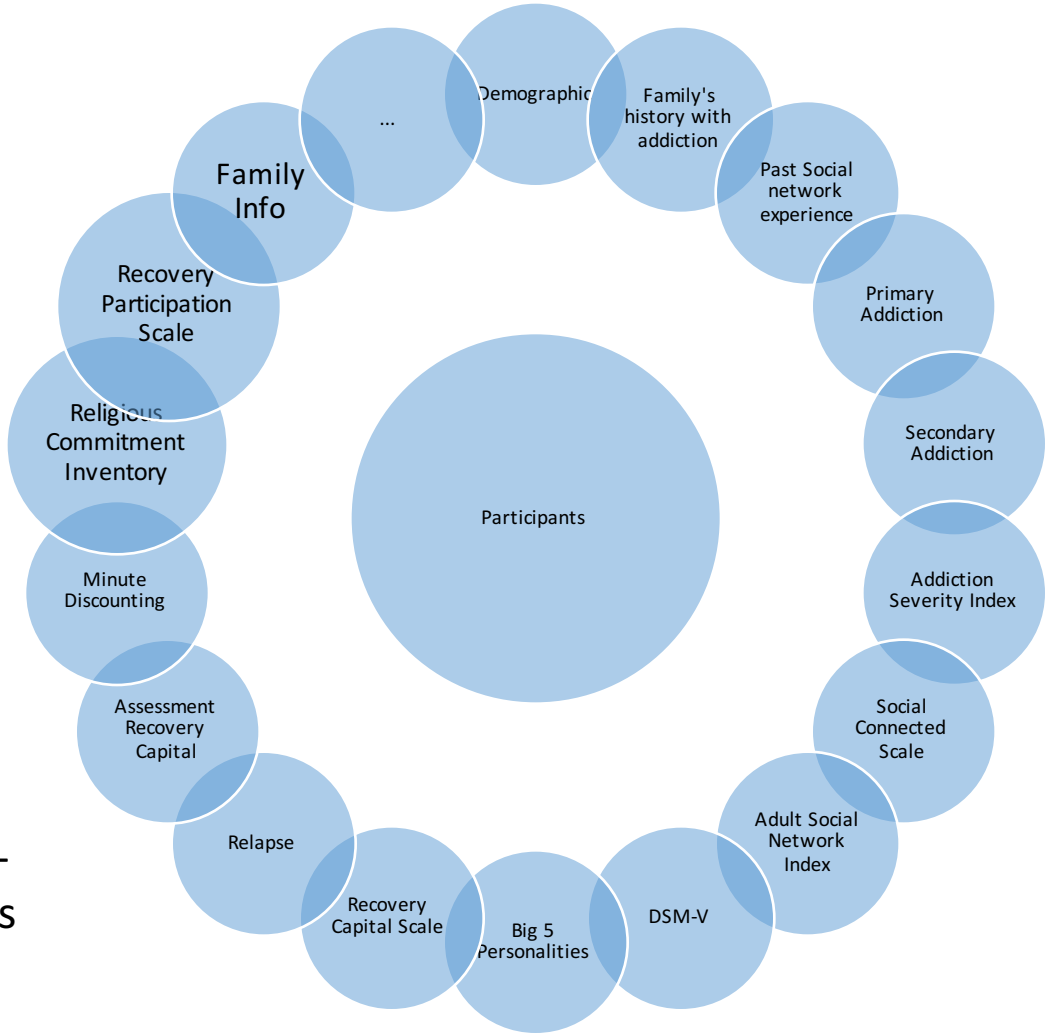
Background: Social Interactome (SI)

- Social Interactome
  - NIH-funded project conducted by a team of researchers
  - Study the community of people, who are recovering from addiction
  - Study their interactions in an online social network, built to provide support and management of their recovery
- The project is broken down into set of “test vs. control” experiments with variables defined:
  - Duration of study
  - Number of participants required
  - Avenue of recruitment
  - Null and alternative hypotheses

Background: SI Setup

- The project is broken down into a set of clinical trials. For each clinical trial:
  - The team decides on a set of null and alternative hypotheses and the duration of the trial
  - Recruits participants for the trial
  - Organizes the participants into one of two (or more) 128-node social network
  - Participants interact with the website and their assigned friends
- Two 16-week clinical trials have been completed. Along with a set of smaller scaled trials executed via Amazon Mturk

Background: SI Participant Info



- Collected from 19,070 questions
- ~10 psychology-based measures
- 16 surveys

Background:SI Website Use Data

Overall Challenges

- How do you organize the data?
- How do you validate/clean the data?
- What do you analyze first? And in what order do you go about it?
- How do you make sense of the data?
- How to interpret psychology-related measures?
- Big goal: *How to streamline the entire process from data collection to analyses to presentation such that it is reproducible and extensible?*

Goal

- Goal: Investigate/explore ways to model the data and recommend an approach.

- Approaches to understand the data
  - Frequency Distributions / Histograms
  - Time series
  - Checking for correlations
  - Comparing means and standard deviations
    - t-tests
  - Statistical modeling

Approaches

- Statistical Modeling
  - What do we model?
    - Substance relapse
    - Engagement/Change in engagement
    - Change in psychology-related measures
    - Change in behavior
    - Homophily
    - Friendship or Trust
  - Factors
    - Classification: What would be the predictor variables? Response variables?
    - PGMs: Directed or Undirected? What would be the factors?

Approaches

- Classification
  - Network-Classification using NetKit-SRL (Statistical Relational Learning)[1] [Focus of the presentation]

- Learning using Markov Logic Networks[2]

1 Sofus A. Macskassy , Foster Provost. "Classification in Networked Data: A toolkit and a univariate case study," Journal of Machine Learning, 8(May):935-983, 2007.

2 Domingos, Pedro and Richardson, Matthew (2007). Markov Logic: A Unifying Framework for Statistical Relational Learning. In L. Getoor and B. Taskar (eds.), Introduction to Statistical Relational Learning (pp. 339-371), 2007. Cambridge, MA: MIT Press.

Network Classification

- Idea: Taking advantage of relational information in addition to attribute information for entity classification. Example: Networked data.
- Focuses on *within-network* classification
- Networks of web pages, research papers, social networks, etc.
- Netkit-SRL: Toolkit developed to employ statistical relational learning and inference

Network Classification

- Netkit-SRL
  - Network learning toolkit for classification and inference
  - Developed by Dr. Macskassy & Dr. Provost
  - Has 3 components
    - Non-relational model
    - Relational model
    - Collective inference
  - Specific Outcomes:
    - Maximize $P(x|G^K)$, where x are labels to be estimated and G^K is everything known in the network
    - Estimating joint distribution over the labels
  - Input:
    - Graph with edges describing relationships and attributes of nodes

Network Classification

- Netkit-SRL Components

	Purpose	Approaches
Local (Non-relational) Classifier	Returns a model which uses only attributes of a node to estimate its class label.	1) Uniform prior; 2) Class-prior
Relational Classifier	Returns a model which uses not only the local attributes of a node but also attributes of related nodes, including their (estimated) class membership.	1) Weighted-vote relational neighbor; 2) Class-distributional relational neighbor; 3) Network-only multinomial Bayes classifier with Markov Random Field estimation
Collective Inference	This module applies collective inference in order to (approximately) maximize the joint probability of the labels of all nodes in the graph whose labels were initially unknown.	1) Relaxation labeling; 2) Iterative classification; 3) Gibb's sampling

Network Classification

- Possible instantiations

Author	Non-relational Classifier	Relational Classifier	Collective Inference
Chakrabarti et al. (1998)[1]	Naïve Bayes classifier	Naïve Bayes Markov Random Field	Relaxation labeling
Lu & Getoor (2003)[2]	Logistic regression	Logistic regression	Iterative classification
Macskassy & Provost (2003)[3]	Classes priors	Majority vote of neighboring classes	Relaxation labeling

[1] Chakrabarti, S., Dom, B., & Indyk, P. (1998). Enhanced Hypertext Categorization Using Hyperlinks. Proceedings of the ACM SIGMOD International Conference on Management of Data (pp. 307–318).
[2] Lu, Q., & Getoor, L. (2003). Link-Based Classification. International Conference on Machine Learning, ICML-2003 (pp. 496–503).
[3] Macskassy, S. A., & Provost, F. (2003). A Simple Relational Classifier. Proceedings of the Second Workshop on Multi-Relational Data Mining (MRDM-2003) at KDD-2003 (pp. 64–76).

Network Classification

- Weighted-vote relational neighbor classifier (wv-RN)
  - Authors: Macskassy, S. A., & Provost, F. (2003)
  - Estimates class membership by assuming existence of homophily
  - Weighted mean of class-membership probabilities of entities in D_e (where D_e is the neighbors of entity/node e)
  - $P(c|e) = \frac{1}{Z}\sum w(e, e_j) * P(c|e_j)$

Network Classification

- Collective Inference using Relaxation Labeling
- Definition of collective inference:

> Given graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{X})$ where X_i is the (single) attribute of vertex $v_i \in \mathbf{V}$, and given known values x_i of X_i for some subset of vertices $\mathbf{V}^K$, *univariate collective inferencing* is the process of simultaneously inferring the values x_i of X_i for the remaining vertices, $\mathbf{V}^U = \mathbf{V} - \mathbf{V}^K$, or a probability distribution over those values.

- Similar but different to Gibbs sampling in that:
  - Keeps track of class probability estimates for X^U
  - Instead of updating the graph one node at a time, updates class probabilities of all vertices, at iteration t+1, based on estimations from step t.

Network Classification: Experiment

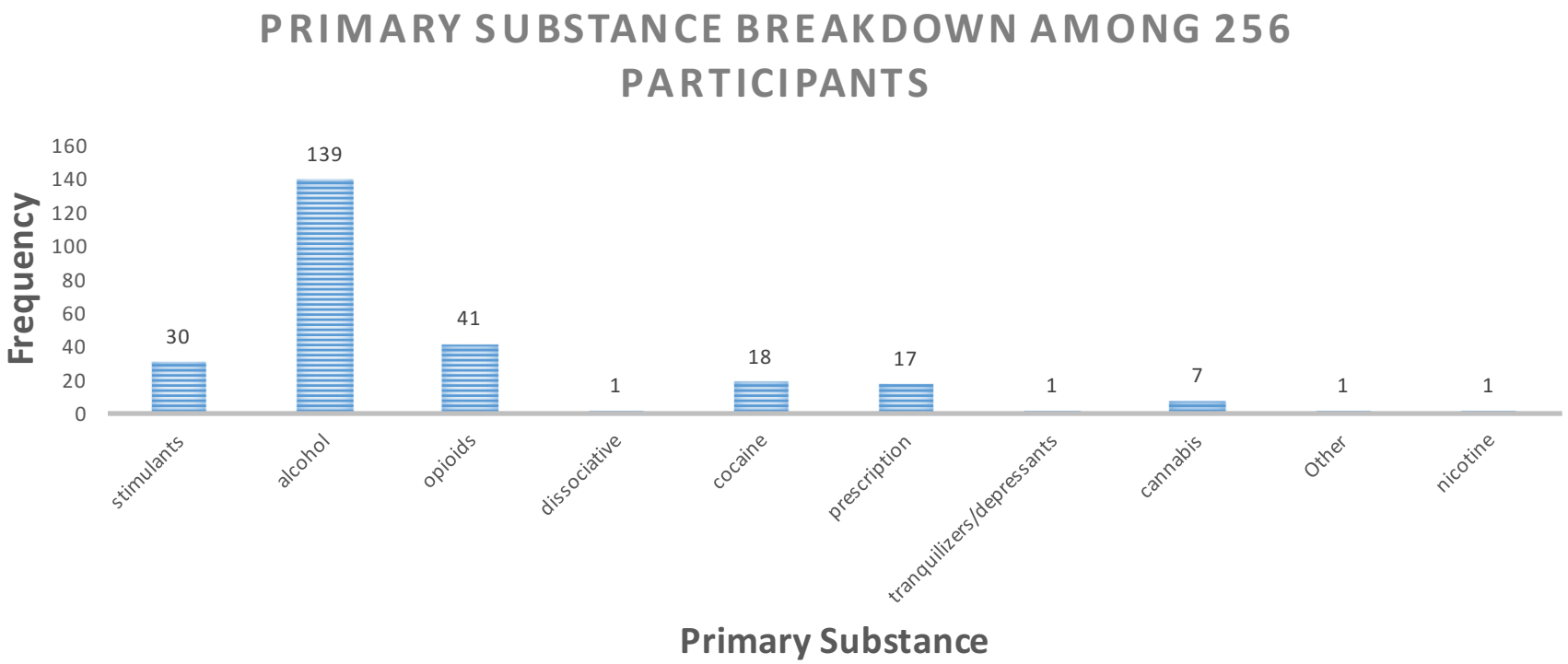
- Experiment
  - Rationale: Participants who are “homopholous” (who have shared background in common), have common interests.
  - Hypothesis: Given a set of common interests, between pairs of participants, one can predict the homophily-measures with good accuracy.
  - Input graph
    - Nodes: Participants
    - Attributes: Addiction, Education, Income
    - Edges: Edge weight is the number of news stories + success stories + educational modules that both nodes (connected via the edge) have viewed in common.
    - Predicted attribute: Addiction

Network Classification: Experiment Config

- Possible Experiment configurations
  - Non-relational classifier: *None*
  - Relational classifier: (Options)
    - Weighted Vote Relational Neighbor
    - Class-Distributional Relational Neighbor
  - Collective inference: (Options)
    - Relaxation labeling
    - Gibbs sampling
    - Iterative classification
  - Data: Nodes and edges extracted from experiment 1 replicate 2 (E1R2) participant interactions.
  - *Goal: Predict 1) Primary Addiction (given graph); 2) Education (given graph); 3) Income bracket (given graph)*

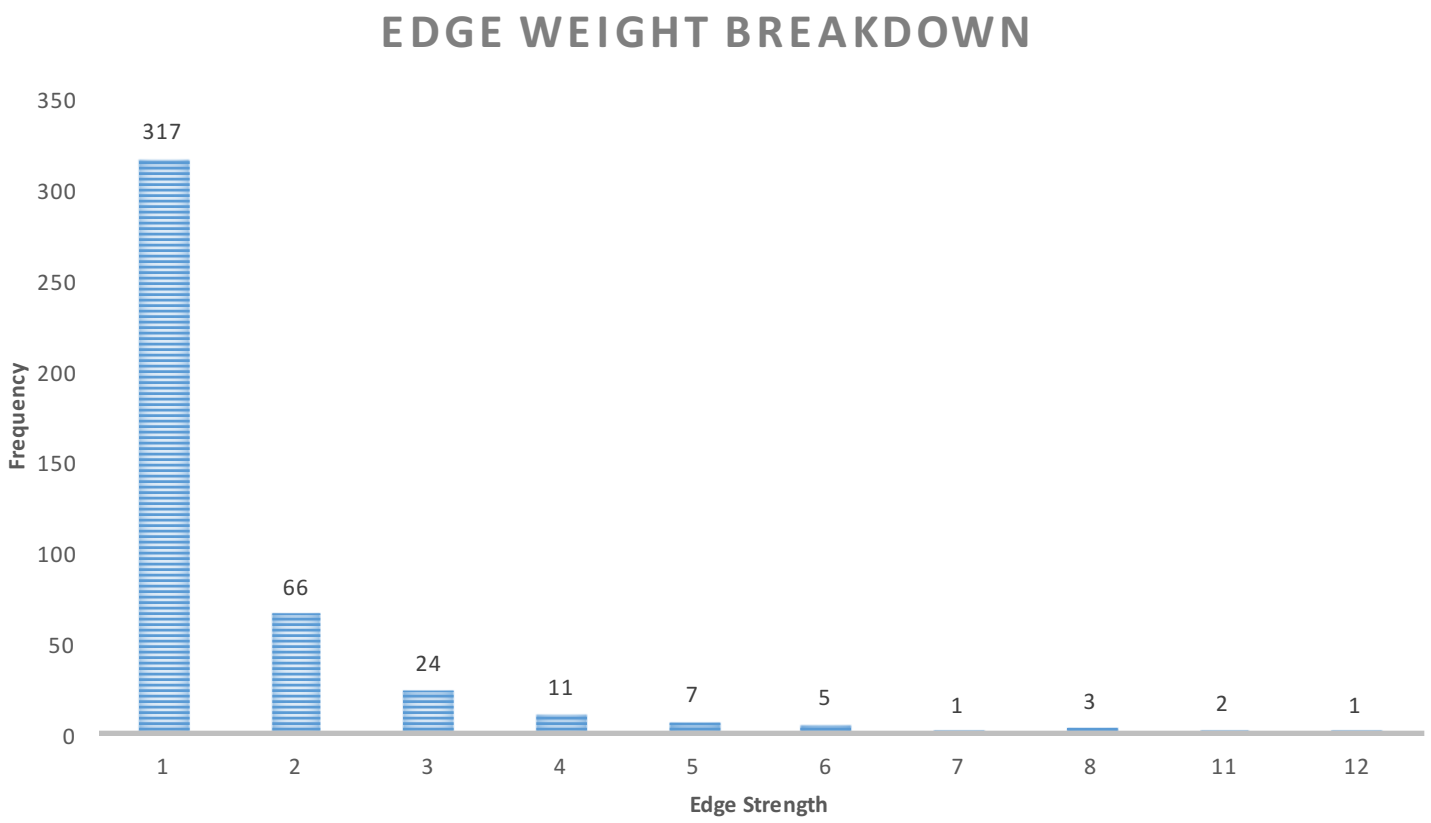
Network Classification: Experiment

- E1R2 data statistics
  - # of nodes: 256; # of edges: 436



Network Classification: Experiment

- Experiment E1R2 data statistics
  - Edge weight breakdown



Network Classification: Experiment Results

- Network classification framework results (various experiment configurations given as row/column names) (Metric: Accuracy)
  - Goal: Predict "Primary Addiction" of participants

Relational Classifier/Collective Inference methods	Relaxation Labeling	Gibbs Sampling	Iterative Classification
Weighted Vote Relational Neighbor (wvRN)	0.36601	0.37908	**0.39216**
Class-Distributional Relational Neighbor	0.15686	0.22222	0.18954

Network Classification: Experiment Results

- Predicted Response/Class = Primary Addiction
- Configuration: wvRN with relaxation labeling
- Confusion Matrix

	00	01	02	03	04	05	06	07	08	09	
stimulants 00:	1	8	1	0	2	1	0	0	1	0	: (1 correct of 14) (accuracy: 0.07143)
alcohol 01:	3	51	16	0	5	1	1	5	1	0	: (51 correct of 83) (accuracy: 0.61446)
opioids 02:	3	14	2	0	2	0	0	2	0	0	: (2 correct of 23) (accuracy: 0.08696)
dissociative 03:	0	0	0	0	0	0	0	0	0	0	: (0 correct of 0) (accuracy: 0)
cocaine 04:	1	10	1	0	1	0	0	1	0	0	: (1 correct of 14) (accuracy: 0.07143)
prescription 05:	1	3	3	0	0	1	0	1	1	0	: (1 correct of 10) (accuracy: 0.1)
tranquilizers 06:	0	0	1	0	0	0	0	0	0	0	: (0 correct of 1) (accuracy: 0)
cannabis 07:	1	2	1	0	2	0	0	0	0	0	: (0 correct of 6) (accuracy: 0)
Other 08:	0	1	0	0	0	0	0	0	0	0	: (0 correct of 1) (accuracy: 0)
nicotine 09:	0	0	1	0	0	0	0	0	0	0	: (0 correct of 1) (accuracy: 0)
TOTAL:	10	89	26	0	12	3	1	9	3	0	: (56 correct of 153) (accuracy: 0.36601)

Network Classification: Experiment Results

- Predicted Response/Class = Education
- Configuration: wvRN with relaxation labeling
- Confusion Matrix

	00	01	02	03	04	05	06	
highschool 00:	0	0	0	0	0	0	0	: (0 correct of 0) (accuracy: 0)
bachelors 01:	0	14	11	0	4	5	2	: (14 correct of 36) (accuracy: 0.38889)
other 02:	0	21	17	0	10	6	2	: (17 correct of 56) (accuracy: 0.30357)
phd 03:	0	2	1	0	1	0	0	: (0 correct of 4) (accuracy: 0)
masters 04:	0	7	8	2	2	3	1	: (2 correct of 23) (accuracy: 0.08696)
diploma 05:	0	7	6	0	0	0	0	: (0 correct of 13) (accuracy: 0)
associates 06:	0	8	8	1	0	2	2	: (2 correct of 21) (accuracy: 0.09524)
TOTAL:	0	59	51	3	17	16	7	: (35 correct of 153) (accuracy: 0.22876)

Network Classification: Experiment Results

- Predicted Response/Class = Income
- Configuration: wvRN with relaxation labeling
- Confusion Matrix

	00	01	02	03	04	05	06	
lessthan30 00:	31	7	5	1	5	1	5	: (31 correct of 55) (accuracy: 0.56364)
lessthan150 01:	14	0	0	0	2	0	1	: (0 correct of 17) (accuracy: 0)
lessthan90 02:	7	0	3	0	3	2	1	: (3 correct of 16) (accuracy: 0.1875)
other 03:	2	0	0	0	2	0	0	: (0 correct of 4) (accuracy: 0)
lessthan50 04:	9	0	2	0	13	2	2	: (13 correct of 28) (accuracy: 0.46429)
above150 05:	6	3	0	0	1	0	0	: (0 correct of 10) (accuracy: 0)
lessthan7 06:	11	1	1	0	5	0	5	: (5 correct of 23) (accuracy: 0.21739)
TOTAL:	80	11	11	1	31	5	14	: (52 correct of 153) (accuracy: 0.33987)

Network Classification: Experiment Conclusion

- Conclusion
  - The highest accuracy for all experiment configurations for predicting primary addiction as shown in slide 22, is 0.392
  - The confusion matrix for predicting each of primary addiction, education and income shows more details on the accuracy of predicting each class.
  - The accuracy is low.
  - This is *probably* due to the fact that our experiment configuration does NOT include a non-relational component.
  - Furthermore, our graph edges, and attributes have only 1-3 fields. The graph needs to be more dense with a lot more information to be used for network-based inference.

Network Classification: Next Steps

- Possible extensions of the work:
  - Build graph with different representation of edges
  - Construct more attributes of the node for non-relational (local) classifier step
  - Try experiments with priors learnt from various traditional classification models.
- Problem/Challenge
  - Extension or further work is open-ended.
  - Part of doctoral work: Build a logical flowchart of inquiries/hypotheses.
  - The logical flowchart of inquiries can be used and called upon based on user's line of inquiry.

Learning via Markov Logic Networks

- A Markov Logic Network (MLN) is a set of pairs (F, w) where
  - F is a formula in first-order logic
  - w is a real number
- Together with a set of constants, it defines a Markov network with
  - One node for each grounding of each predicate in the MLN
  - One feature for each grounding of each formula F in the MLN, with the corresponding weight w

*Slide source: http://www.cs.washington.edu/homes/pedrod/psrai.ppt

Learning via Markov Logic Networks

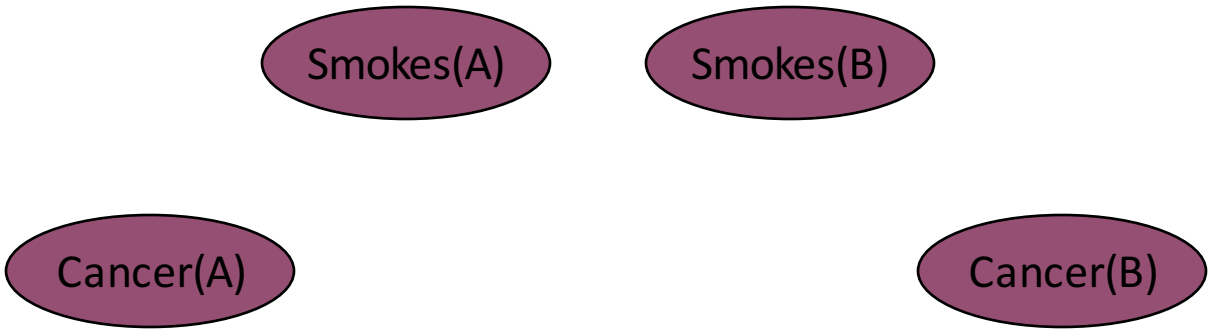
Smoking causes cancer.

Friends have similar smoking habits.

$$\forall x\ Smokes(x) \Rightarrow Cancer(x)$$

$$\forall x, y\ Friends(x, y) \Rightarrow \left(Smokes(x) \Leftrightarrow Smokes(y)\right)$$

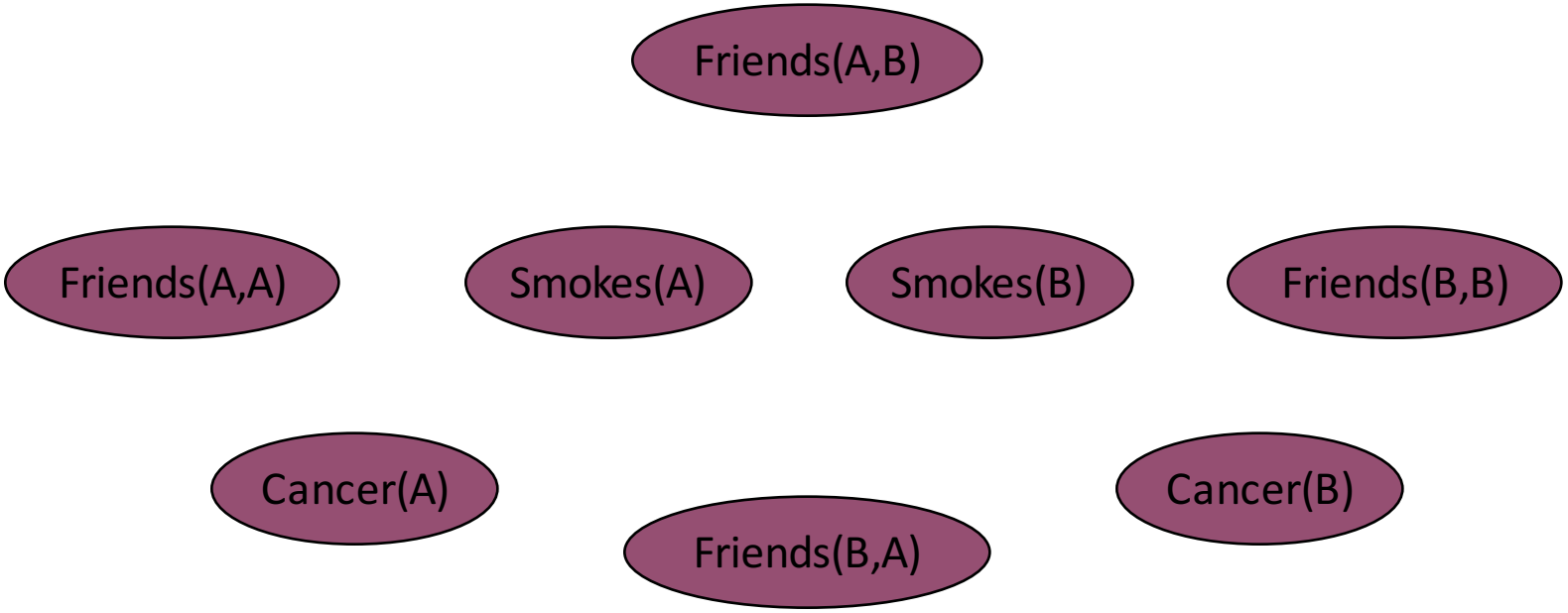
Two constants: **Anna** (A) and **Bob** (B)



*Slide source: http://www.cs.washington.edu/homes/pedrod/psrai.ppt

Learning via Markov Logic Networks

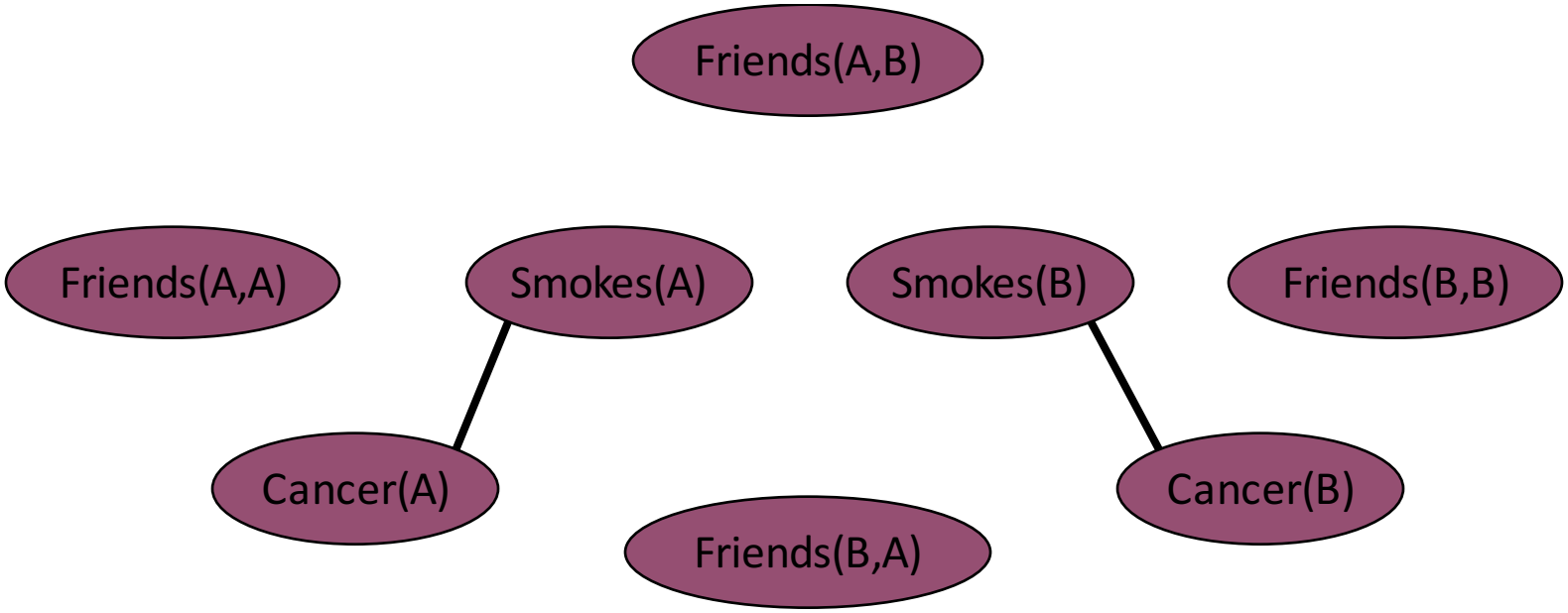
1.5	$\forall x\ Smokes(x) \Rightarrow Cancer(x)$
1.1	$\forall x, y\ Friends(x, y) \Rightarrow \left(Smokes(x) \Leftrightarrow Smokes(y)\right)$



*Slide source: http://www.cs.washington.edu/homes/pedrod/psrai.ppt

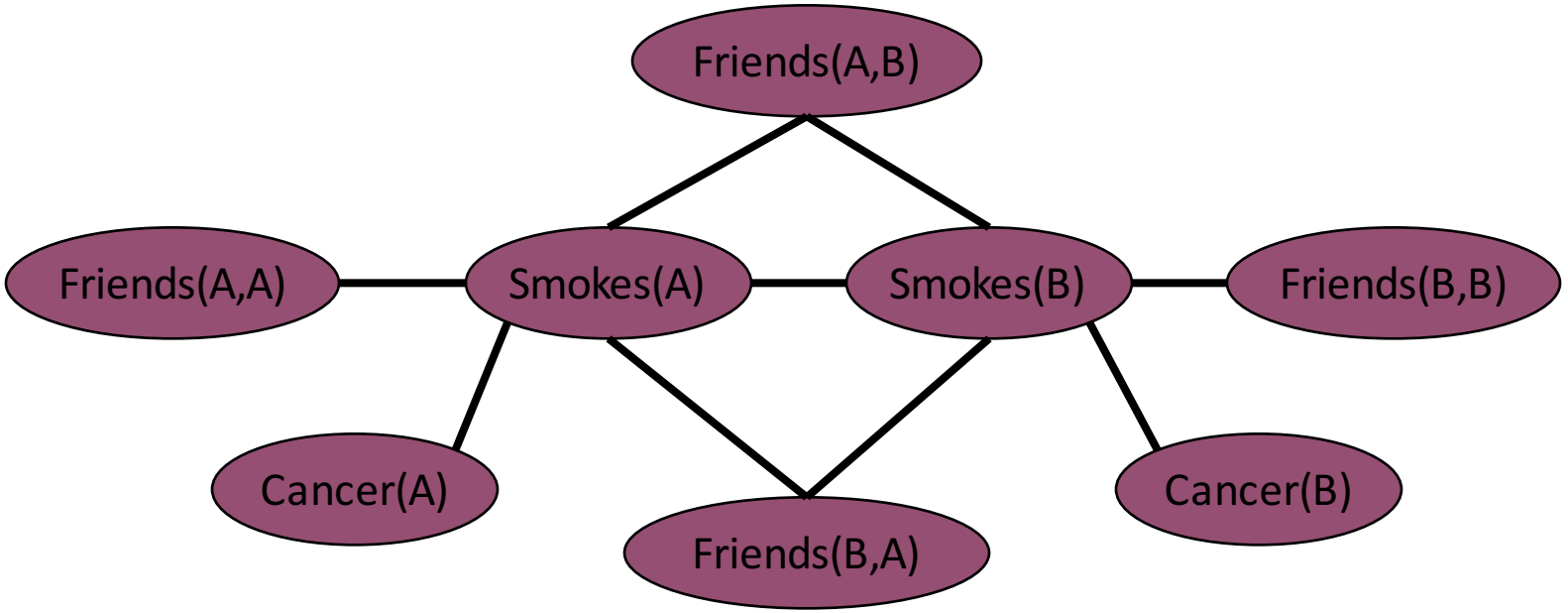
Learning via Markov Logic Networks

1.5	$\forall x\ Smokes(x) \Rightarrow Cancer(x)$
1.1	$\forall x, y\ Friends(x, y) \Rightarrow \left(Smokes(x) \Leftrightarrow Smokes(y)\right)$



*Slide source: http://www.cs.washington.edu/homes/pedrod/psrai.ppt

Learning via Markov Logic Networks



$$\text{Probability of a world } x\text{: } P(x) = \frac{1}{Z}\exp\left(\sum_i w_i n_i(x)\right)$$

Weight of formula *i*

No. of true groundings of formula *i* in *x*

*Slide source: http://www.cs.washington.edu/homes/pedrod/psrai.ppt

Learning via Markov Logic Networks

Tasks/Applications

- Basics
- Logistic regression
- Hypertext classification
- Information retrieval
- Entity resolution
- Hidden Markov models
- Information extraction
- Statistical parsing
- Semantic processing
- Bayesian networks
- Relational models
- Robot mapping
- Planning and MDPs
- Practical tips

*Slide source: http://www.cs.washington.edu/homes/pedrod/psrai.ppt

Future work

- Next steps
  - Extract more attributes for each participant
  - Compiledifferent ways to represent edge weight
  - Build local classifier and testing results for Netkit-SRL
  - Use Alchemy to represent data using Markov Logic networks.

Questions?

Network Classification

- Other works
  - Inductive logic programming
  - Markov random fields
  - Conditional random fields
  - Probabilistic relational models
  - Relational Bayesian networks
  - Relational dependency networks
  - Relational Markov networks