Probabilistic Models of Networks and other Relational Data

Zoubin Ghahramani

Department of Engineering
University of Cambridge, UK

zoubin@eng.cam.ac.uk
http://mlg.eng.cam.ac.uk/

Cambridge Network Day
2014


Konstantina Palla


David Knowles


Creighton Heaukulani

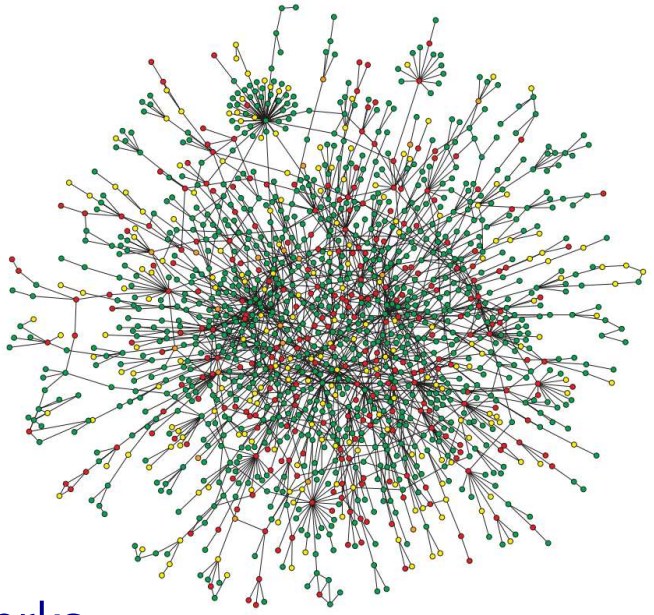

James Lloyd


Peter Orbanz


Dan Roy

Modelling Networks

We are interested in modelling networks.



Biological networks: protein-protein interaction networks

Social networks: friendship networks; co-authorship networks

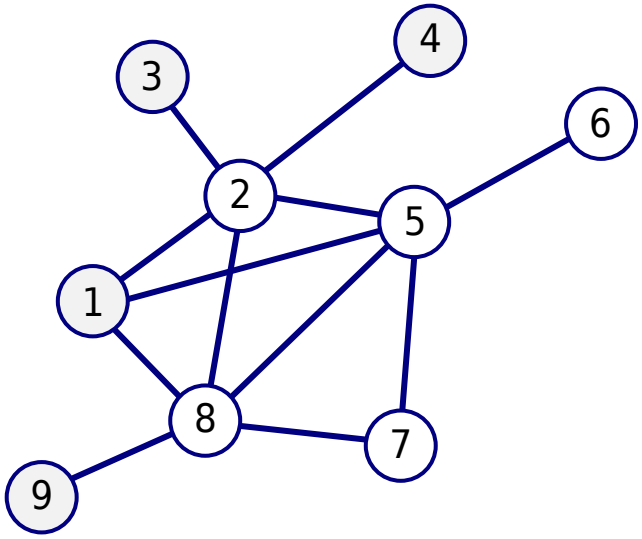
We wish to have models that will be able to

- predict missing links,
- infer latent properties or classes of the objects,
- generalise learned properties from smaller observed networks to larger networks.

Figure from Barabasi and Oltvai 2004: A protein-protein interaction network of budding yeast

What is a network?

- A set $\mathcal{V}$ of **entities** (nodes, vertices) and
- A set $\mathcal{Y}$ of pairwise **relations** (links, edges) between the entities



We can represent this as a graph with a binary adjacency matrix $\mathbf{Y}$ where element $y_{ij} = 1$ represents a link between nodes v_i and v_j

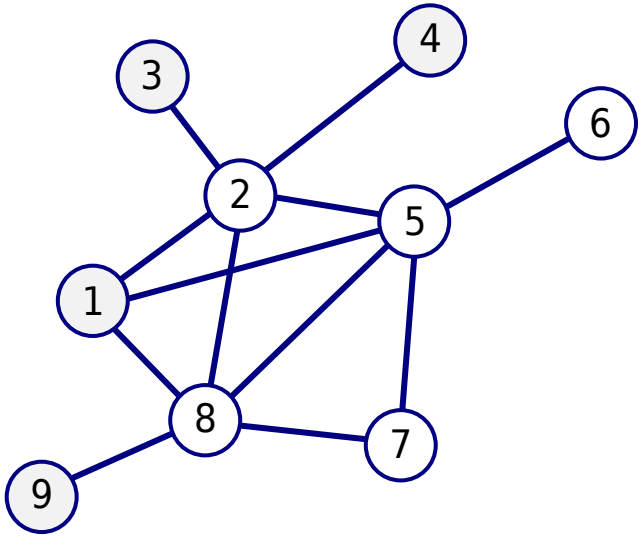
We'll focus on undirected graphs (i.e. networks of symmetric relations) but much of what is discussed extends to more general graphs.

What is a model?

Descriptive statistics: identify interesting properties of a network (e.g. degree distribution)

Predictive or generative model: A model that could generate random networks and predict missing links, etc.

Erdös-Rényi Model

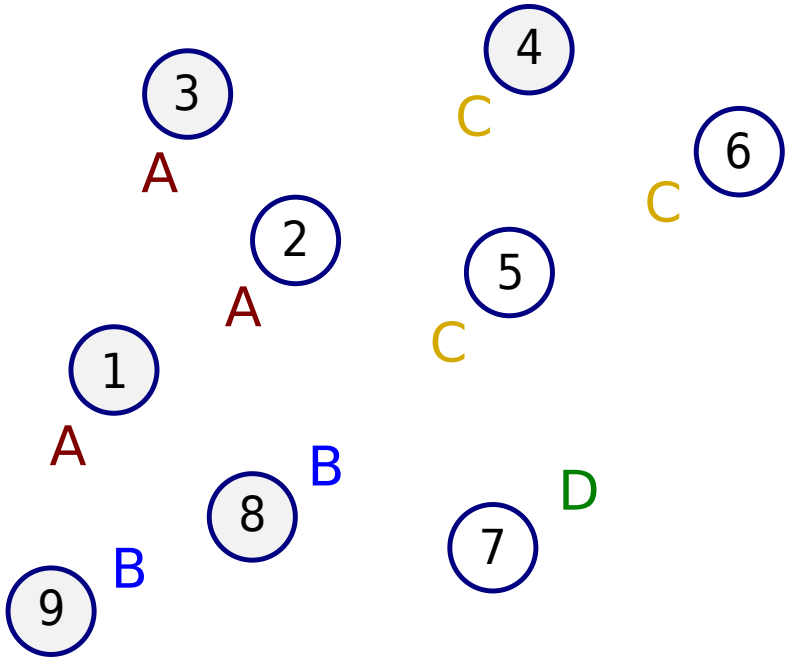


A very simple model that assumes each link is independent, and present with probability $\pi \in [0, 1]$

$$y_{ij} \sim \text{Bern}(\pi)$$

This model is easy to analyse but does not have any interesting structure or make any nontrivial predictions. The only thing one can learn from such a model is the average density of the network.

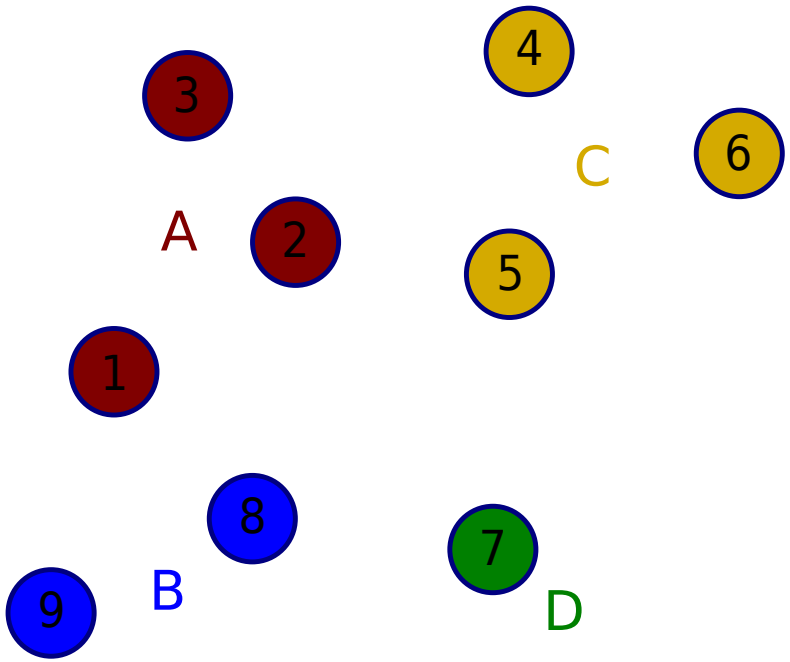
Latent Class Models



The basic idea is to posit that the structure of the network arises from latent (or hidden) variables associated with each node.

We can think of latent class models as having a single discrete hidden variable associated with each node.

Latent Class Models

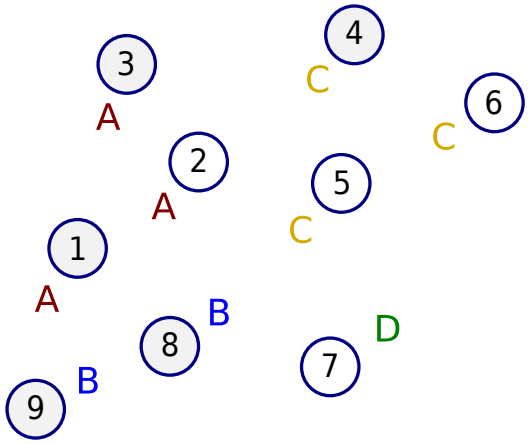


This corresponds to a *clustering* of the nodes.
Such models can be used for *community detection*.

For example, the discrete hidden variables might correspond to the political views of each individual in a social network.

Latent Class Models

Stochastic Block Model (Nowicki and Snijders, 2001)



Each node v_i has a hidden class from a set of K possible classes: $c_i \in \{1, \ldots, K\}$

For all i:

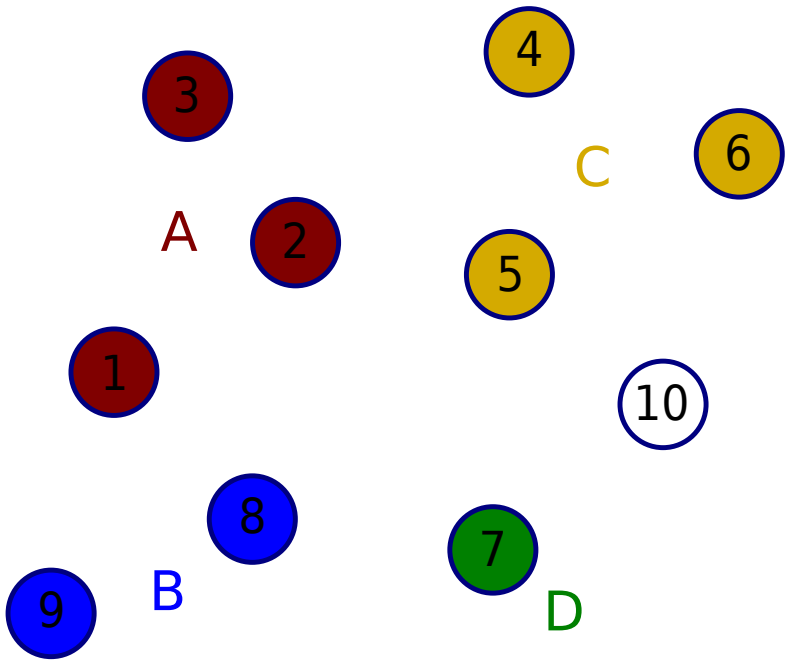
$$c_i \sim \text{Discrete}(p_1, \ldots p_K)$$

The probability of a link between two nodes v_i and v_j depends on their classes:

$$P(y_{ij} = 1 | c_i = k, c_j = \ell) = \rho_{k\ell}$$

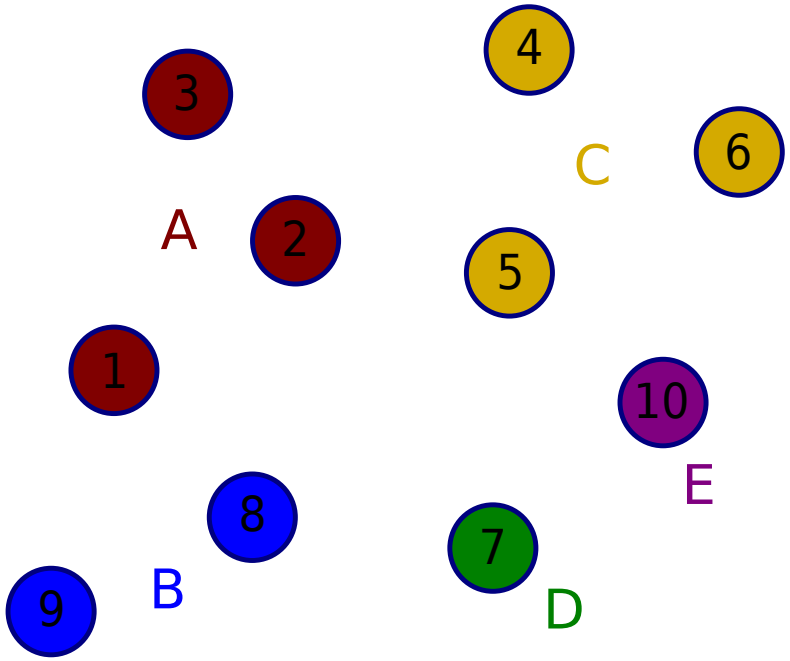
The parameters of the model are the $K \times 1$ class proportion vector $\mathbf{p} = (p_1, \ldots, p_K)$ and the $K \times K$ link probabillity matrix $\boldsymbol{\rho}$ where $\rho_{k\ell} \in [0, 1]$.

Latent Class Models



If we observe a new node, which class do we assign it to?

Nonparametric Latent Class Models



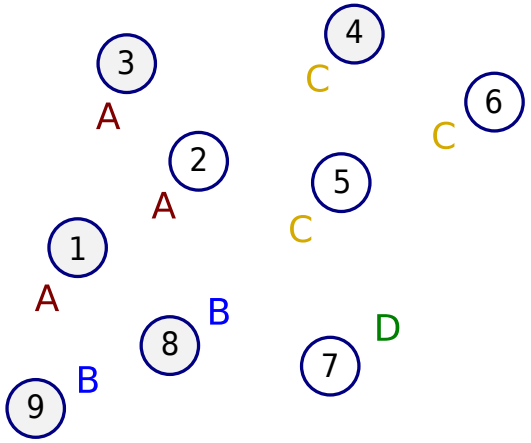
The new node could belong to one of the previously observed classes, but might also belong to an as yet unobserved class.

This motivates *nonparametric* models, where the number of observed classes can grow with the number of nodes.[1]

[1] Nonparametric models are sometimes called *infinite* models since they allow infinitely many classes, features, parameters, etc.

Nonparametric Latent Class Models

Infinite Relational Model (Kemp et al 2006)



Each node v_i has a hidden class $c_i \in \{1, \ldots, \infty\}$

For all i:

$$c_i | c_1, \ldots, c_{i-1} \sim \mathrm{CRP}(\alpha)$$ [2]

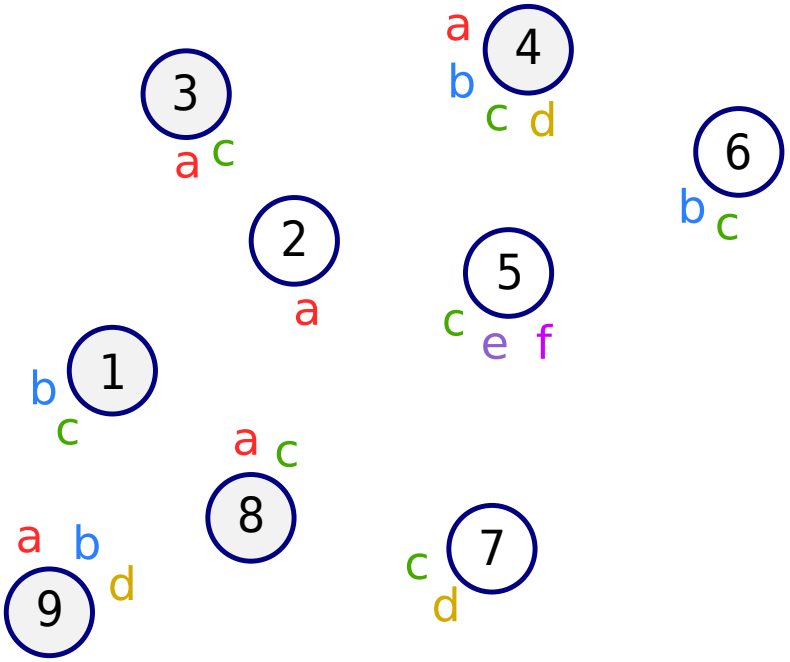
As before, probability of a link between two nodes v_i and v_j depends on their classes:

$$P(y_{ij} = 1 | c_i = k, c_j = \ell) = \rho_{k\ell}$$

Note that $\boldsymbol{\rho}$ is an infinitely large matrix, but if we give each element a beta prior we can integrate it out. Inference done via MCMC. Fairly straightforward to implement.

[2] CRP, or *Chinese Restaurant Process*, is an exchangeable distribution on partitions of the integers which is used to define clustering models with an unbounded number of clusters.

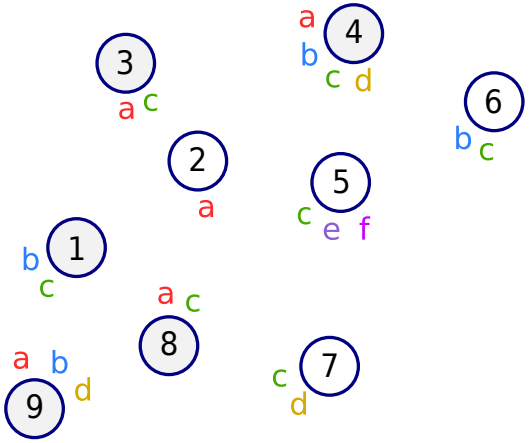
Latent Feature Models



- Each node posses some number of latent features.
- Alternatively we can think of this model as capturing *overlapping clusters or communities*
- The link probability depends on the latent features of the two nodes.
- The model should be able to accommodate a potentially unbounded (infinite) number of latent features.

Latent Feature Models

Nonparametric Latent Feature Relational Model (Miller et al 2010)



Let $z_{ik} = 1$ denote whether node i has feature k

The latent binary matrix $\mathbf{Z}$ is drawn from an IBP[3] distribution: $\mathbf{Z}|\alpha \sim \text{IBP}(\alpha)$

The elements of the parameter matrix $\mathbf{W}$ are drawn iid from: $w_{k\ell} \sim \text{N}(0, \sigma^2)$

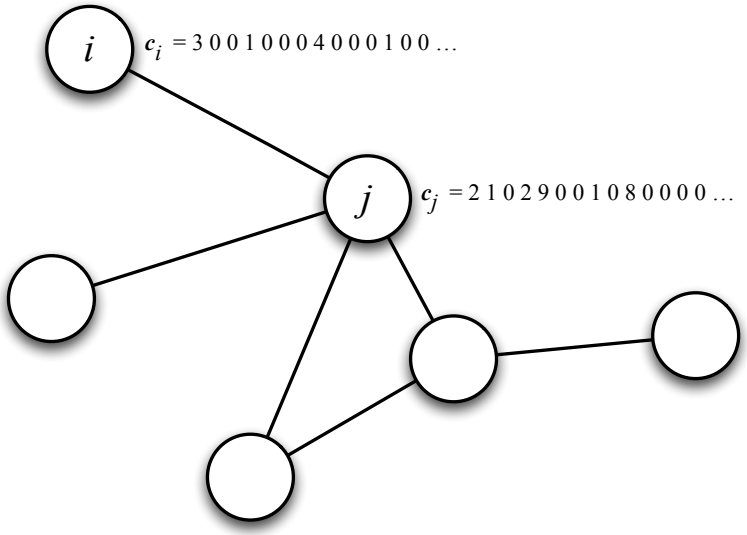
The link probability is:

$$P(y_{ij} = 1|\mathbf{W}, \mathbf{Z}) = \sigma\left(\sum_{k,\ell} z_{ik}\ z_{j\ell}\ w_{k\ell}\right)$$

where $\sigma(\cdot)$ is the logistic (sigmoid) function.

[3] An IBP, or *Indian Buffet Process*, is an exchangeable distribution over infinite feature allocations—generalising clustering models to allow overlapping communities.

Infinite Latent Attribute model for network data



- Each object has some number of latent attributes
- Each attribute can have some number of discrete values
- Probability of a link between object i and j depends on the attributes of i and j:

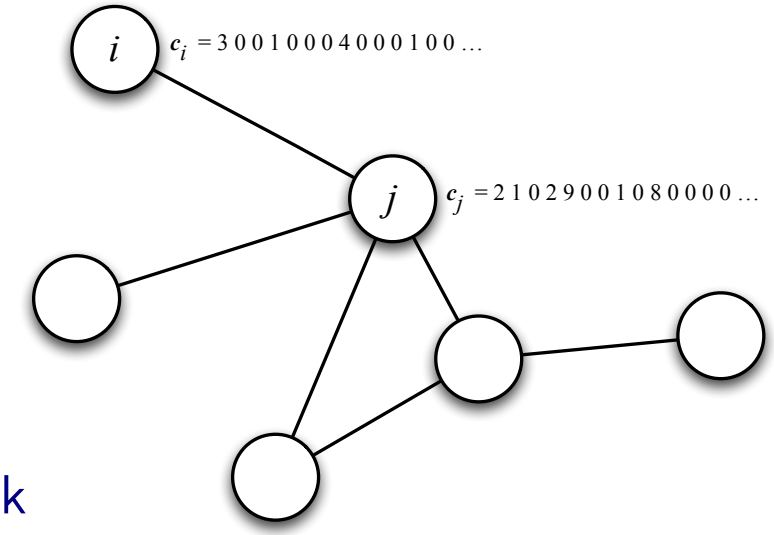
$$P(y_{ij} = 1 | \mathbf{z}_i, \mathbf{z}_j, \mathbf{C}, \mathbf{W}) = \sigma\Big(\sum_m z_{im} z_{jm} w^{(m)}_{c_i^m c_j^m} + s \Big)$$

- Potentially unbounded number of attributes, and values per attribute[4]
- Generalises both the IRM and the NLFRM.

(w/ Konstantina Palla, David Knowles, ICML 2012)

[4] An IBP is used for the attribute matrix, $\mathbf{Z}$ and a CRP for the values of each attribute, $\mathbf{C}$

Infinite Latent Attribute model for network data



Example: a student friendship network

- Each student might be involved in some activities or have some features:
  `person_i` has attributes `(College, sport, politics)`
  `person_j` has attributes `(College, politics, religion, music)`

- Each attribute has some values:
  `person_i` = `(College=Trinity, sport=squash, politics=LibDem)`
  `person_j` = `(College=Kings, politics=LibDem, religion=Catholic, music=choir)`

- Prob. of link between person i and j depends on their attributes and values.
- The attributes and values are *not observed*—they are learned from the network.

Infinite Latent Attribute: Results

Test error rates (missing link prediction) on NIPS coauthorship, and gene interaction prediction benchmark datasets.

	IRM	LFIRM	ILA
NIPS	0.0440 ± 0.0014	0.0228 ± 0.0041	$\mathbf{0.0106 \pm 0.0007}$
Genes	0.3608 ± 0.0031	0.2661 ± 0.0086	$\mathbf{0.0735 \pm 0.0047}$

IRM: (Kemp and Tenenbaum 2006)
LFIRM: (Miller, Griffiths and Jordan 2010)

Exchangeable Sequences

Exchangeable sequence:
A sequence is exchangeable if its joint distribution is invariant under arbitrary permutation of the indices:

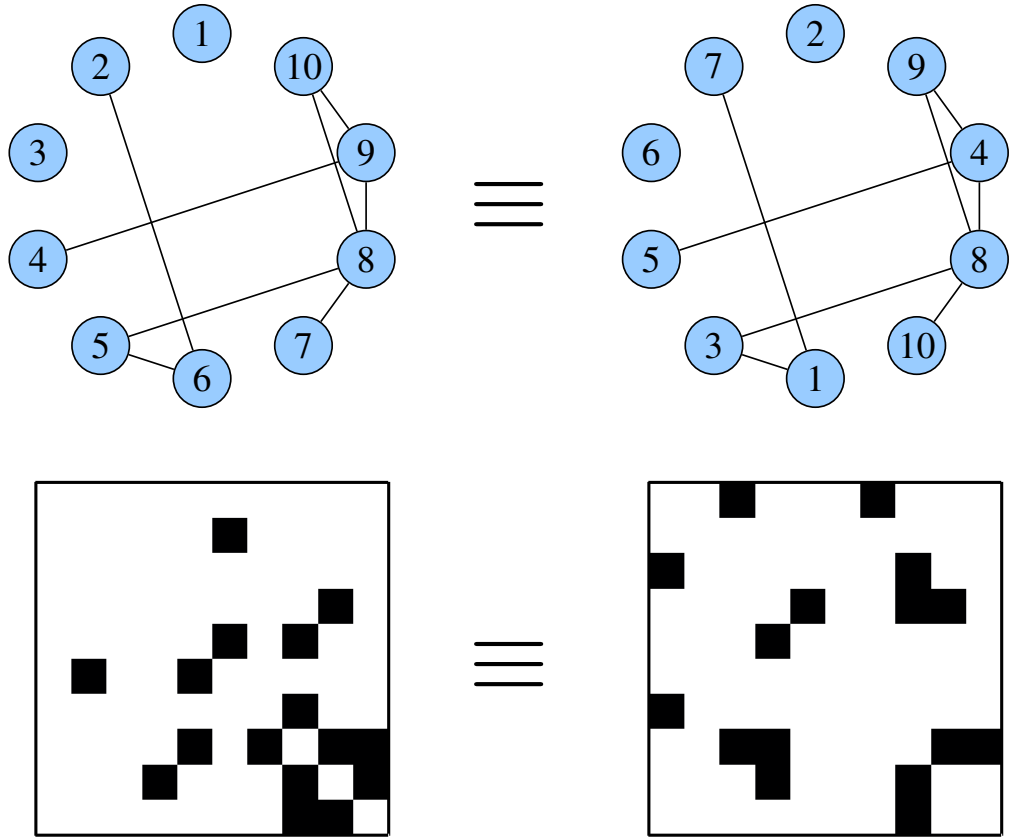
$$(X_1, X_2, ...) \stackrel{d}{=} (X_{\pi(1)}, X_{\pi(2)}, ...) \ \ \forall \pi \in S_\infty.$$

de Finetti's Theorem:
$(X_i)_{i \in N}$ is exchangeable if and only if there exists a random probability measure Θ on X such that $X_1, X_2, \ldots | \Theta \ \sim \ \text{iid} \ \Theta$

Interpretation:
Any probabilistic model of data which assumes that the order of the data does not matter, can be expressed as a Bayesian mixture of iid models. Note that Θ may in general need to be infinite dimensional (i.e. *nonparameteric*).

Exchangeable Arrays

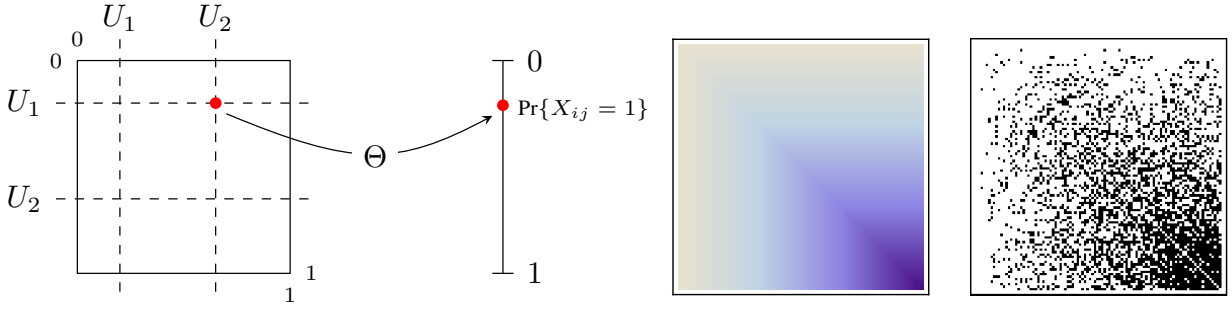
Exchangeable arrays: An array $X = (X_{ij})_{i,j\in\mathbb{N}}$ is called an exchangeable array if $(X_{ij}) \stackrel{d}{=} (X_{\pi(i)\pi(j)})$ for every $\pi \in S_\infty$.



Exchangeable Arrays

Exchangeable arrays: An array $X = (X_{ij})_{i,j\in\mathbb{N}}$ is called an exchangeable array if $(X_{ij}) \stackrel{d}{=} (X_{\pi(i)\pi(j)})$ for every $\pi \in S_\infty$.

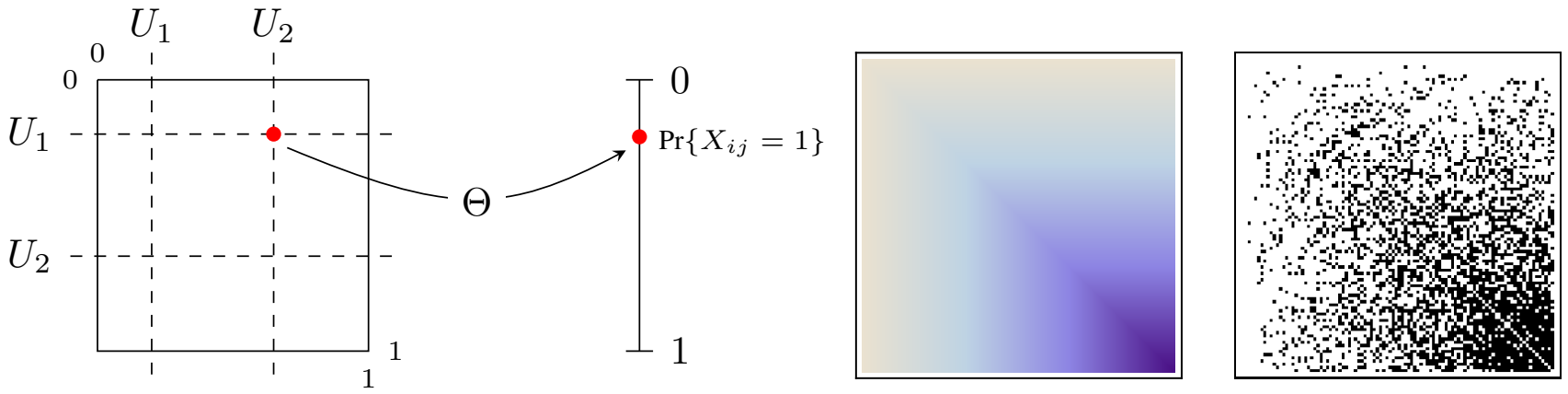
Aldous-Hoover Theorem:
A random matrix (X_{ij}) is exchangeable if and only if there is a random (measurable) function $F : [0,1]^3 \to X$ such that $(X_{ij}) \stackrel{d}{=} (F(U_i, U_j, U_{ij}))$ for every collection $(U_i)_{i\in\mathbb{N}}$ and $(U_{ij})_{i\leq j\in\mathbb{N}}$ of i.i.d. Uniform$[0,1]$ random variables, where $U_{ji} = U_{ij}$ for $j < i \in \mathbb{N}$.



Interpretation:
Any model of matrices, arrays (or graphs) where the order of rows and columns (nodes) is irrelevant can be expressed by assuming *latent variables* associated with each row and column, and a *random function* mapping these latent variables to the observations.

Random Function Model

We develop a nonparametric probabilistic model for arrays and graphs that makes explicit the Aldous Hoover representation:



$$\Theta \sim \mathrm{GP}(0, \kappa) \tag{1}$$

$$U_1, U_2, \ldots \overset{\text{iid}}{\sim} \mathrm{Uniform}[0, 1] \tag{2}$$

$$W_{ij} = \Theta(U_i, U_j) \tag{3}$$

$$X_{ij} \sim P[\cdot | W_{ij}] \tag{4}$$

(w/ James Lloyd, Dan Roy, Peter Orbanz, NIPS 2012)

Random Function Model

The random function model can be related to a number of existing models for matrices, arrays/tensors, and graphs.

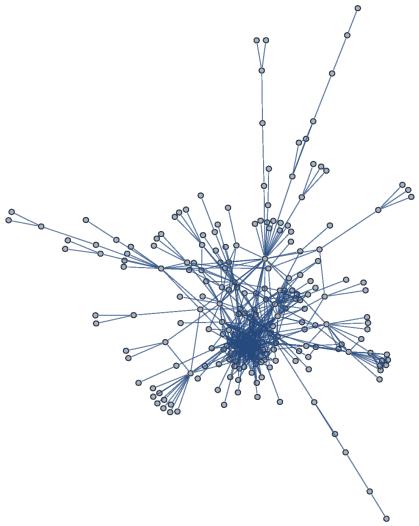
Graph data

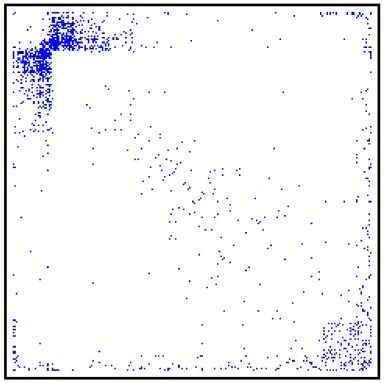
Random function model	Θ	$\sim$	$\mathcal{GP}(0, \kappa)$
Latent class	W_{ij}	$=$	$m_{U_i U_j}$ where $U_i \in \{1, \ldots, K\}$
IRM	W_{ij}	$=$	$m_{U_i U_j}$ where $U_i \in \{1, \ldots, \infty\}$
Latent distance	W_{ij}	$=$	$-\lvert U_i - U_j \rvert$
Eigenmodel	W_{ij}	$=$	$U_i' \Lambda U_j$
LFRM	W_{ij}	$=$	$U_i' \Lambda U_j$ where $U_i \in \{0, 1\}^\infty$
ILA	W_{ij}	$=$	$\sum_d \mathbb{I}_{U_{id}} \mathbb{I}_{U_{jd}} \Lambda^{(d)}_{U_{id} U_{jd}}$ where $U_i \in \{0, \ldots, \infty\}^\infty$
SMGB	Θ	$\sim$	$\mathcal{GP}(0, \kappa_1 \otimes \kappa_2)$

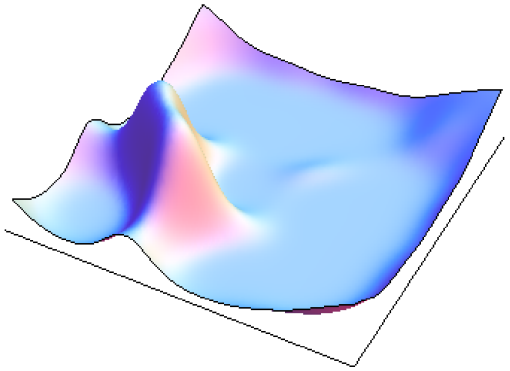
Real-valued array data

Random function model	Θ	$\sim$	$\mathcal{GP}(0, \kappa)$
Mondrian process based	Θ	$=$	piece-wise constant random function
PMF	W_{ij}	$=$	$U_i' V_j$
GPLVM	Θ	$\sim$	$\mathcal{GP}(0, \kappa \otimes \delta)$

Random Function Model: Results







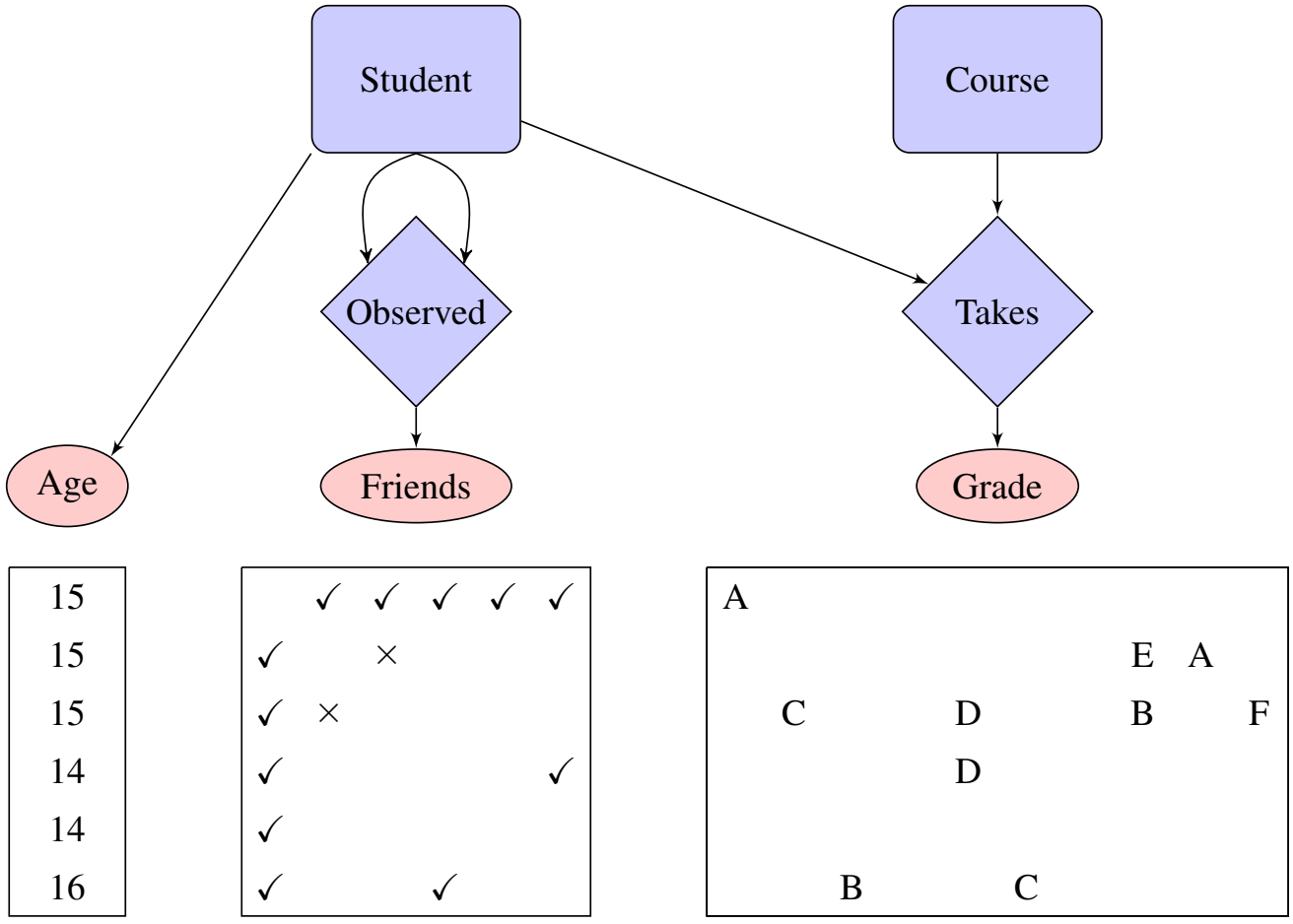
AUC results

Data set	High school			NIPS			Protein		
Latent dimensions	1	2	3	1	2	3	1	2	3
PMF	0.747	0.792	0.792	0.729	0.789	0.820	0.787	0.810	0.841
Eigenmodel	0.742	0.806	0.806	0.789	0.818	0.845	0.805	0.866	0.882
GPLVM	0.744	0.775	0.782	0.888	0.876	0.883	0.877	0.883	0.873
RFM	**0.815**	**0.827**	**0.820**	**0.907**	**0.914**	**0.919**	**0.903**	**0.910**	**0.912**

Relational Data

Networks are special case of *relational data*.

More generally we should think about modelling **databases** containing multiple types of entitites, multiple relations, and features of entities (covariates).



James Lloyd

Do the Aldous-Hoover class of representations extend to such relational data?

Extension: Relational Data wth Covariate Features

Suppose that in addition to a social network (X_{ij}) we have side information in the form of covariates for the users (C_i).

Corollary

Let $(X_{ij})_{i,j\in\mathbb{N}}$ and $(C_i)_{i\in\mathbb{N}}$ be random variables in $\mathcal{X}$ and $\mathcal{X}'$ respectively. Then the following are equivalent:

i. $(X_{ij}), (C_i) \stackrel{d}{=} (X_{p(i)p(j)}), (C_{p(i)})$ *for every* $p \in \mathbb{S}_\infty$.

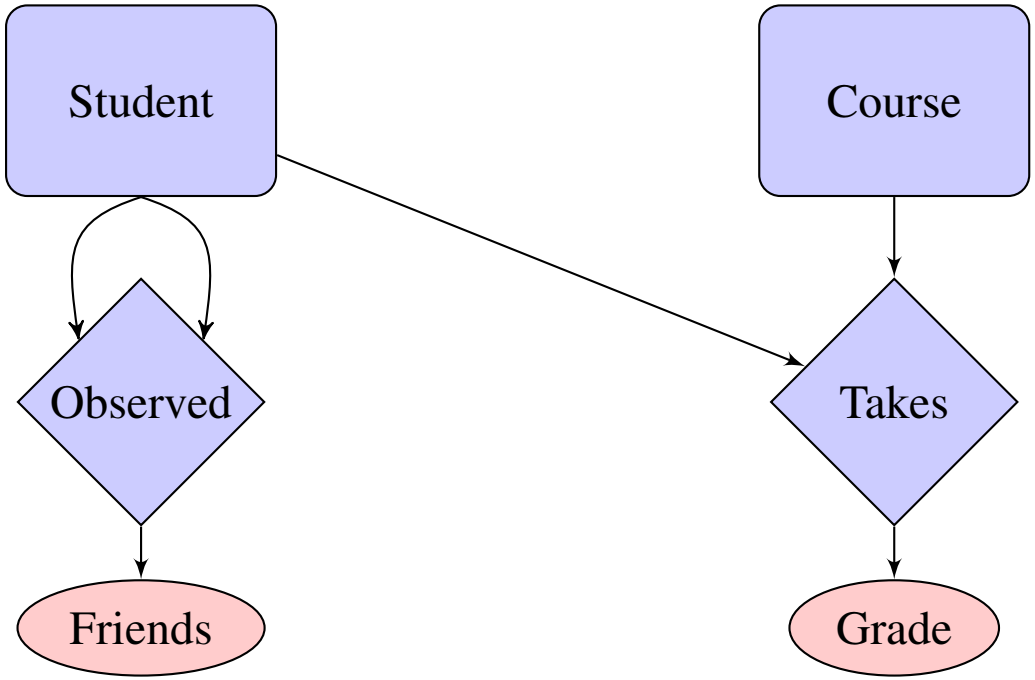
ii. *There are random measurable functions* $F : [0,1]^3 \to \mathcal{X}$ *and* $G : [0,1] \to \mathcal{X}'$ *such that*

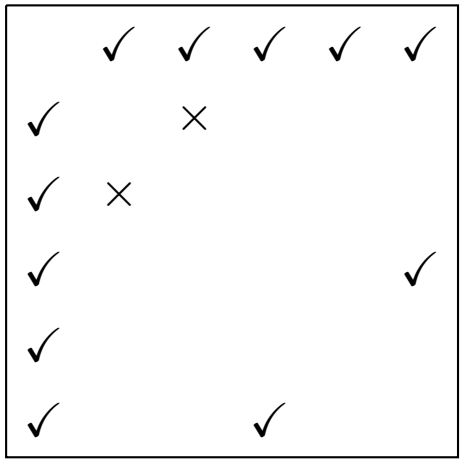
$$(X_{ij}), (C_i) \stackrel{d}{=} (F(U_i, U_j, U_{ij})), (G(U_i)),$$

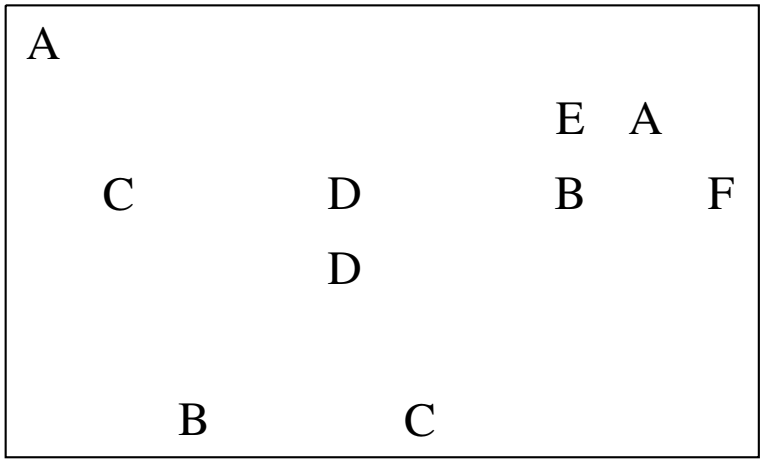
where $(U_i)_{i\in\mathbb{N}}$ *and* $(U_{ij})_{i\leq j\in\mathbb{N}}$ *are i.i.d.* Uniform$[0,1]$ *random variables and* $U_{ji} = U_{ij}$ *for* $j < i \in \mathbb{N}$.

(w/ James Lloyd, Dan Roy, Peter Orbanz)

What about Multiple Relations?







Extension: Two Arrays

Consider rating data (X_{ij}) with users i and items j, and a social network (S_{ik}) over users i, k.

Corollary

The following are equivalent

i. $(X_{ij}), (S_{ik}) \stackrel{d}{=} (X_{p(i)p'(j)}), (S_{p(i)p(k)})$ *for every* $p, p' \in \mathbb{S}_\infty$.

ii. *There exist random functions* F, G *such that*

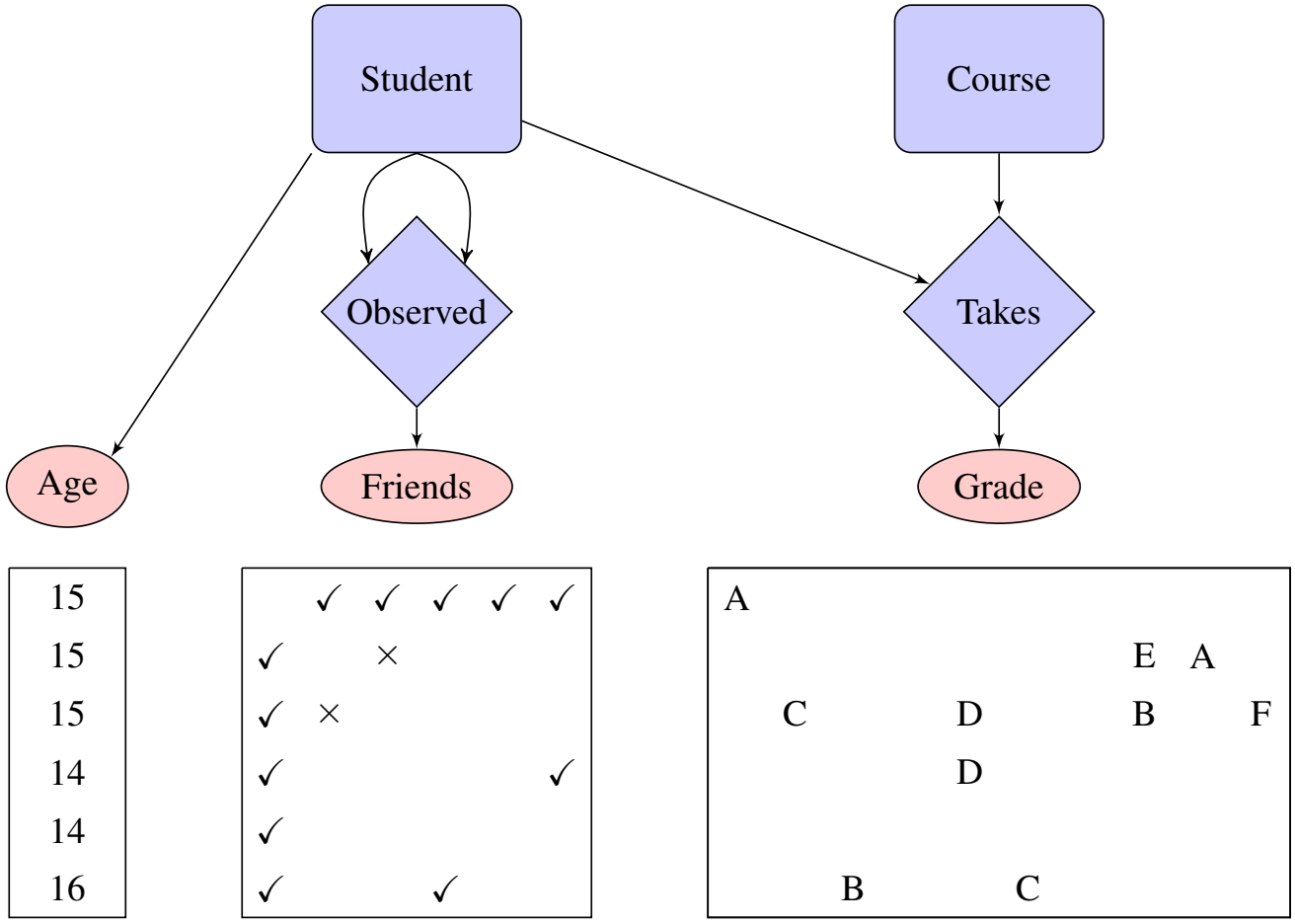
$$(X_{ij}), (S_{ik}) \stackrel{d}{=} (F(U_i, V_j, W_{ij})), (G(U_i, U_k, U_{ik}))$$

where $(U_i)_{i\in\mathbb{N}}, (V_j)_{j\in\mathbb{N}}, (W_{ij})_{i,j\in\mathbb{N}}$ *and* $(U_{ik})_{i\leq k\in\mathbb{N}}$ *are i.i.d.* Uniform$[0, 1]$ *random variables, and* $U_{ki} = U_{ik}$ *for* $k < i \in \mathbb{N}$.

In fact we have extensions to arbitrary databases with R relations and O objects.

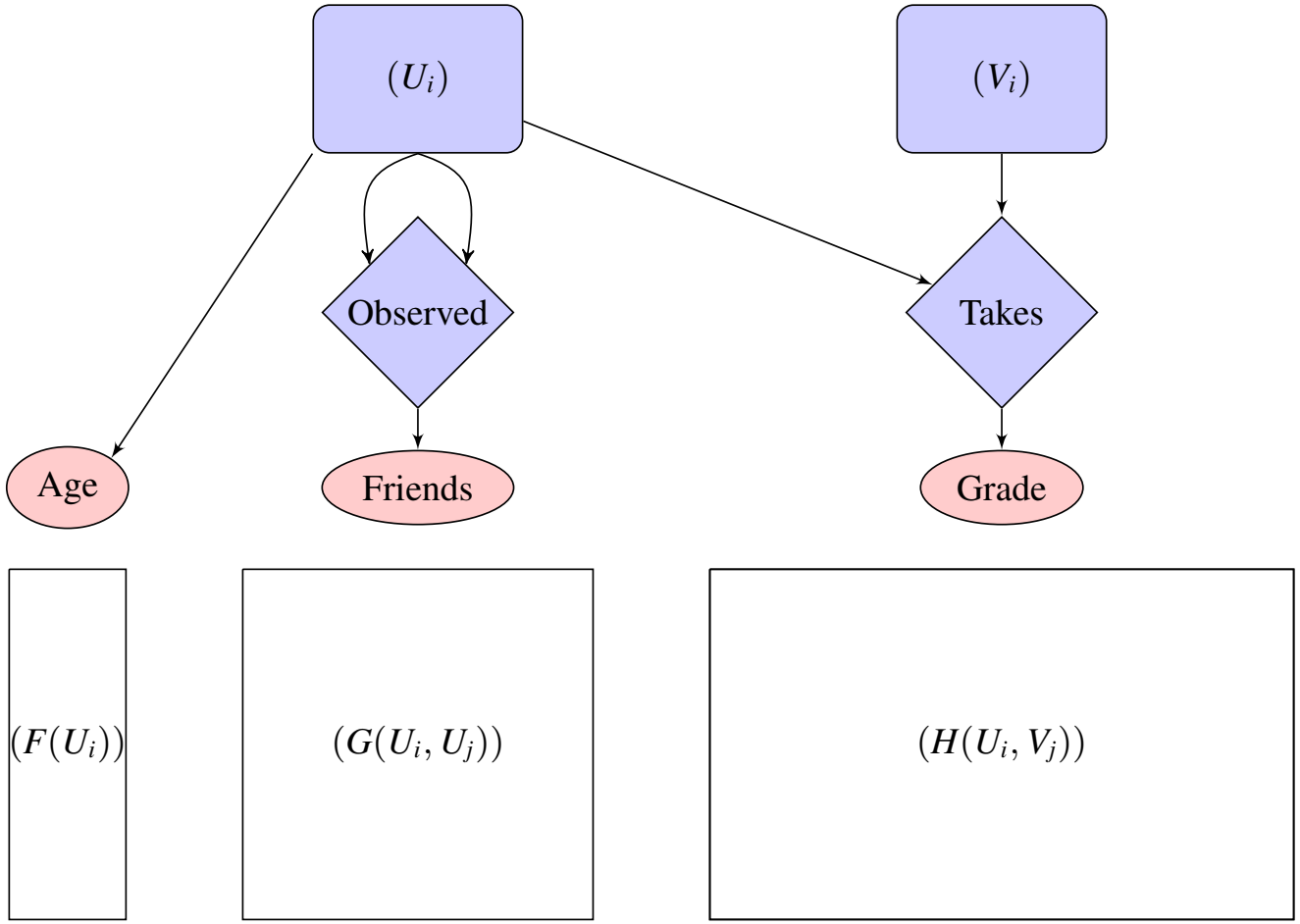
Exchangeable Databases: Summary

To model an exchangeable database...



Exchangeable Databases: Summary

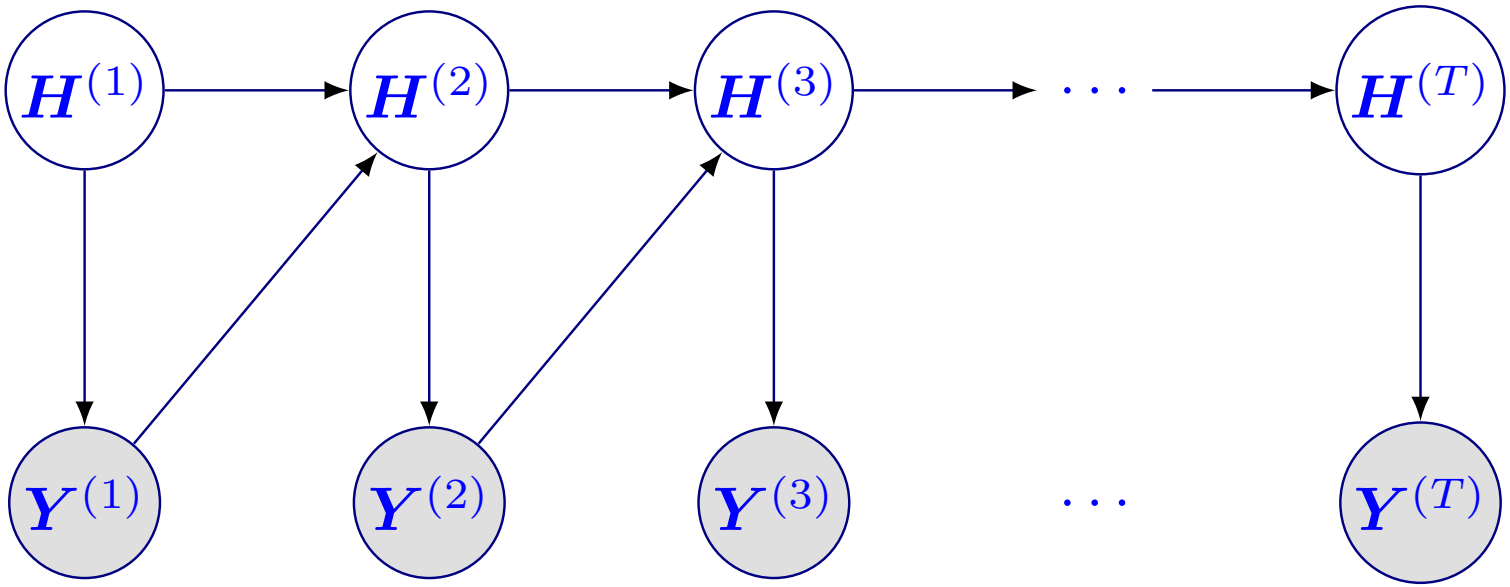
...model each object with a latent variable, and each relation with a random function.



Dynamic Networks

- We observe T slices of a network $(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}, \ldots, \boldsymbol{Y}^{(T)})$
- Assume the structure can be represented by feature matrices $(\boldsymbol{H}^{(1)}, \ldots, \boldsymbol{H}^{(T)})$

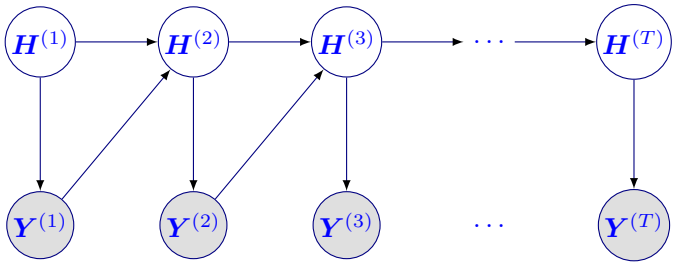
Latent Feature Propagation Models for Dynamic Networks



- Network structure at time t depends on latent features at time t.
- Network stucture at time t influences latent features at time $t+1$: latent feature information propagates between nodes in the network.
- This seems a very intuitive property for a model of social networks.

(w/ Heaukulani, ICML 2013)

Latent Feature Propagation Models for Dynamic Networks



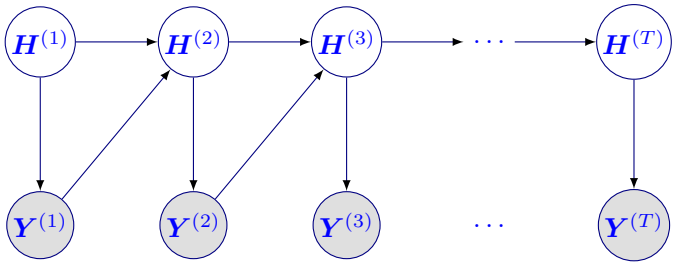
We use the following model

$$h_{ik}^{(t+1)} | \mu_{ik}^{(t+1)} \sim \text{Bernoulli}\left[\sigma\left(c_k\left[\mu_{ik}^{(t+1)} - b_k\right]\right)\right]$$

$$\mu_{ik}^{(t+1)} = (1-\lambda_i) h_{ik}^{(t)} + \lambda_i \frac{h_{ik}^{(t)} + \sum_{j \in \varepsilon(i,t)} w_j h_{jk}^{(t)}}{1 + \sum_{j' \in \varepsilon(i,t)} w_{j'}}$$

1. $\lambda_i \in [0,1]$: a measure of person i's susceptibility to the influence of friends, and $(1-\lambda_i)$ is the corresponding measure of person i's social independence;
2. $w_i \in \mathbb{R}_+$: the weight of influence of person i;
3. $c_k \in \mathbb{R}_+$: a scale parameter for the persistence of feature k;
4. $b_k \in \mathbb{R}_+$: a bias parameter for feature k.

Latent Feature Propagation Models for Dynamic Networks



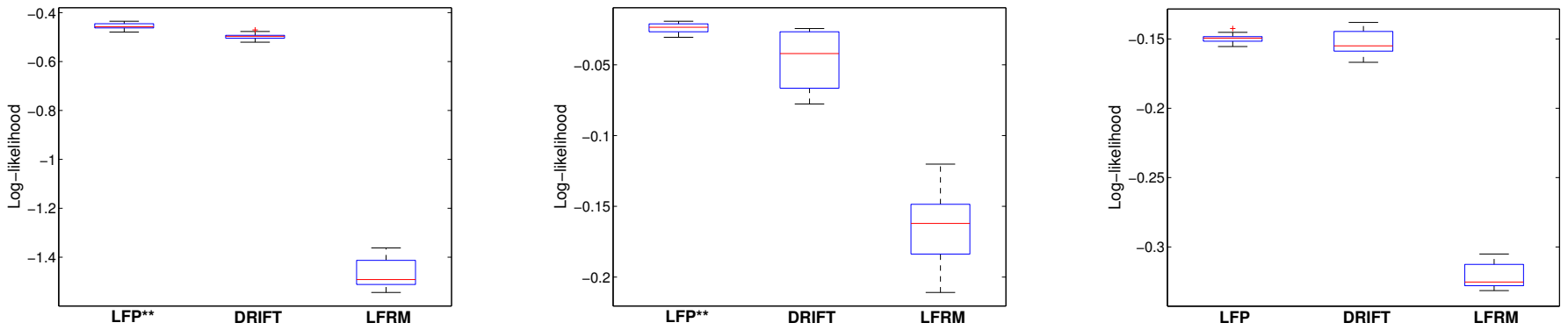
The probability of a link y_{ij} given the latent features $\mathbf{h}_i$ and $\mathbf{h}_j$ is similar to (Miller et al, 2010) LRFRM:

$$y_{ij}^{(t+1)} | \boldsymbol{h}_i^{(t+1)}, \boldsymbol{h}_j^{(t+1)} \sim \text{Bernoulli}(\pi_{ij})$$

$$\pi_{ij} = \sigma\left(\boldsymbol{h}_i^{(t+1)\,T} \boldsymbol{V} \boldsymbol{h}_j^{(t+1)} + s\right)$$

$$\nu_{kk'} \sim \mathcal{N}(0, \sigma_\nu^2).$$

Prediction of Missing Links

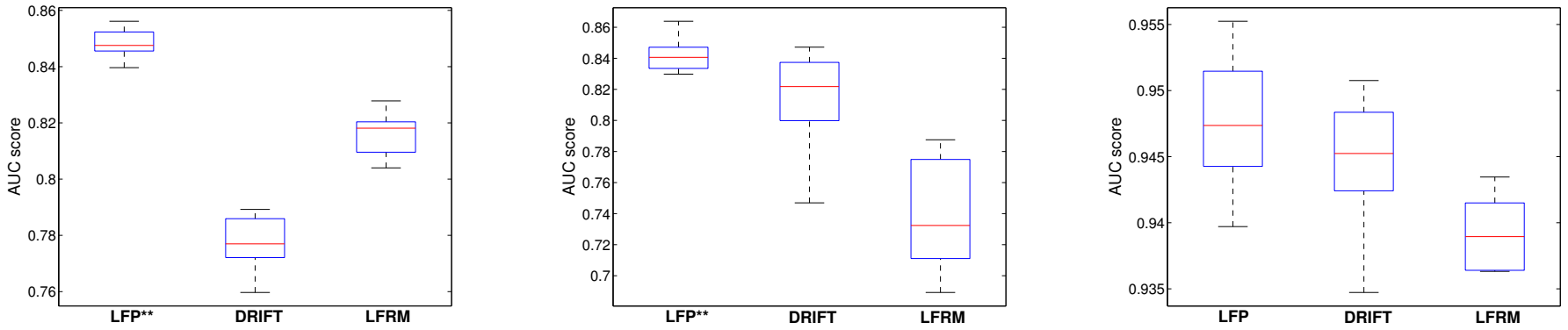


(a) Simulated data (b) NIPS dataset (c) INFOCOM dataset

Figure 1: Log-likelihood of the test edges. Boxplots are over 10 repeats, each holding out a different 20% of the edges. All results are averaged over 300 samples drawn from the steady state distribution following a burn-in period. Statistically significant results are indicated by a $(\star\star)$ based on a T-test at a 0.05 significance level.

DRIFT: (Foulds et al 2011)
LFRM: (Miller et al 2010)

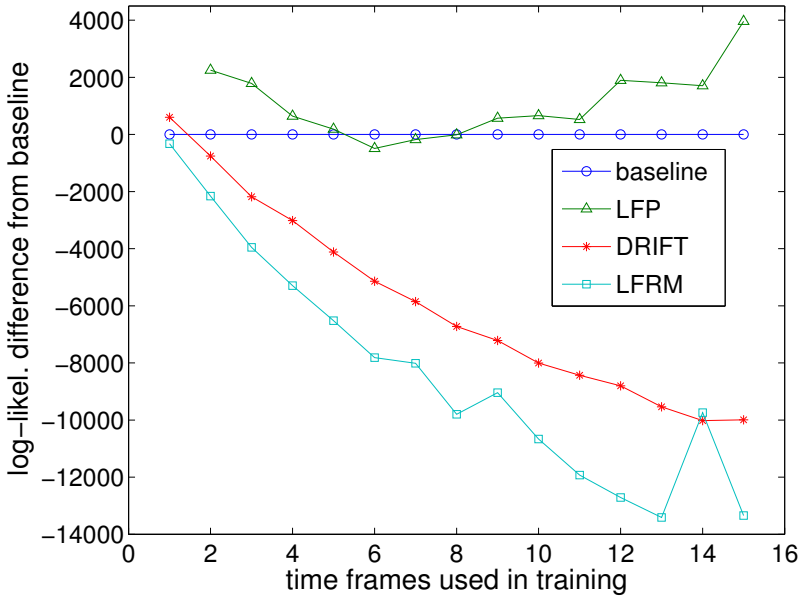
Prediction of Missing Links



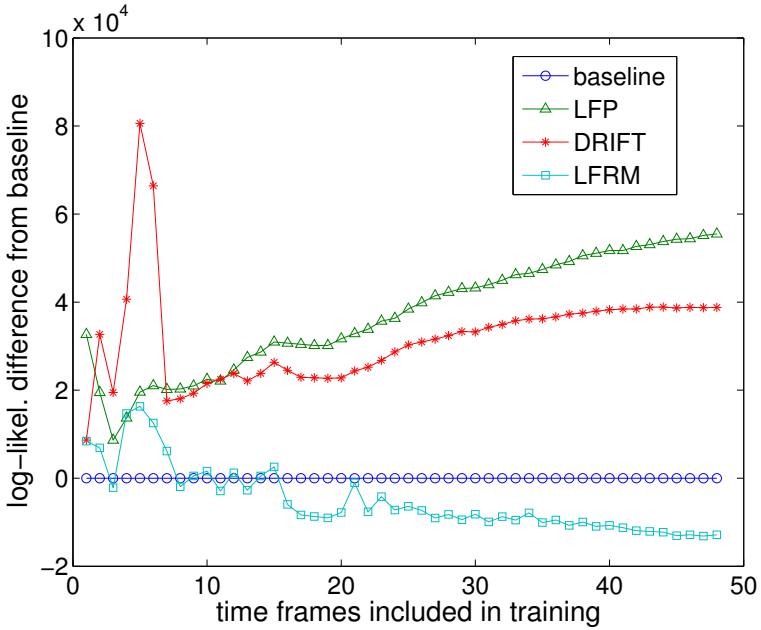
(a) Simulated data (b) NIPS dataset (c) INFOCOM dataset

Figure 2: AUC scores for classifying the test edges in the prediction experiment. Statistically significant results are indicated by $(\star\star)$ based on a T-test at a 0.05 significance level.

Forecasting Future Networks



(a) NIPS dataset, $K = 15$



(b) INFOCOM dataset, $K = 10$

Figure 3: Forecasting a future unseen network. Differences from a naive baseline of the log-likelihoods of $\boldsymbol{Y}^{(t)}$ after training on $\boldsymbol{Y}^{(1:t-1)}$.

Summary

I discussed the general theory of exchangeable arrays, and how this relates to network and relational modelling.

Three network models:

- Infinite Latent Attribute Model
- Random Function Model for Arrays and Relations
- Latent Feature Propagation

Theme: probabilistic models with rich latent variable structures are useful for modelling networks.

Thanks to

Konstantina Palla

David Knowles

Creighton Heaukulani

James Lloyd

Peter Orbanz

Dan Roy

http://learning.eng.cam.ac.uk/zoubin

zoubin@eng.cam.ac.uk

Appendix

Bayesian Machine Learning

Everything follows from two simple rules:

Sum rule: $P(x) = \sum_y P(x, y)$

Product rule: $P(x, y) = P(x)P(y|x)$

$$P(\theta|\mathcal{D}, m) = \frac{P(\mathcal{D}|\theta, m)P(\theta|m)}{P(\mathcal{D}|m)}$$

$P(\mathcal{D}\|\theta, m)$	likelihood of parameters θ in model m
$P(\theta\|m)$	prior probability of θ
$P(\theta\|\mathcal{D}, m)$	posterior of θ given data $\mathcal{D}$

Prediction:

$$P(x|\mathcal{D}, m) = \int P(x|\theta, \mathcal{D}, m)P(\theta|\mathcal{D}, m)d\theta$$

Model Comparison:

$$P(m|\mathcal{D}) = \frac{P(\mathcal{D}|m)P(m)}{P(\mathcal{D})}$$

$$P(\mathcal{D}|m) = \int P(\mathcal{D}|\theta, m)P(\theta|m)\, d\theta$$

Parametric vs Nonparametric Models

- *Parametric models* assume some finite set of parameters θ. Given the parameters, future predictions, x, are independent of the observed data, $\mathcal{D}$:

$$P(x|\theta, \mathcal{D}) = P(x|\theta)$$

  therefore θ capture everything there is to know about the data.

- So the complexity of the model is bounded even if the amount of data is unbounded. This makes them not very flexible.

- *Non-parametric models* assume that the data distribution cannot be defined in terms of such a finite set of parameters. But they can often be defined by assuming an *infinite dimensional* θ. Usually we think of θ as a *function*.

- The amount of information that θ can capture about the data $\mathcal{D}$ can grow as the amount of data grows. This makes them more flexible.

Some References

- Airoldi, E.M., Blei, D.M., Fienberg, S.E., and Xing, E.P. (2008) Mixed-membership stochastic blockmodels. JMLR, 9:19812014.
- Foulds, J., DuBois, C., Asuncion, A.U., Butts, C.T., and Smyth, P. (2011) A dynamic relational infinite fea- ture model for longitudinal social networks. In Proc. AISTATS, April 2011.
- Griffiths, T.L., and Ghahramani, Z. (2011) The Indian buffet process: An introduction and review. *Journal of Machine Learning Research* **12**(Apr):1185–1224.
- Heaukulani, C. and Ghahramani, Z. (2013) Dynamic Probabilistic Models for Latent Feature Propagation in Social Networks. ICML 2013.
- Kemp, C., J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. (2006) Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence*.
- Lloyd, J., Orbanz, P., Ghahramani, Z., Roy, D. (2012) Random function priors for exchangeable arrays with applications to graphs and relational data. NIPS 2012.
- Miller, K.T., T. L. Griffiths, and M. I. Jordan. (2010) Nonparametric latent feature models for link predictions. In *Advances in Neural Information Processing Systems 22*.
- Nowicki, K. and Snijders, T. A. B. (2001) Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96:1077–1087.
- Palla, K., Knowles, D.A., and Ghahramani, Z. (2012) An infinite latent attribute model for network data. In ICML 2012.