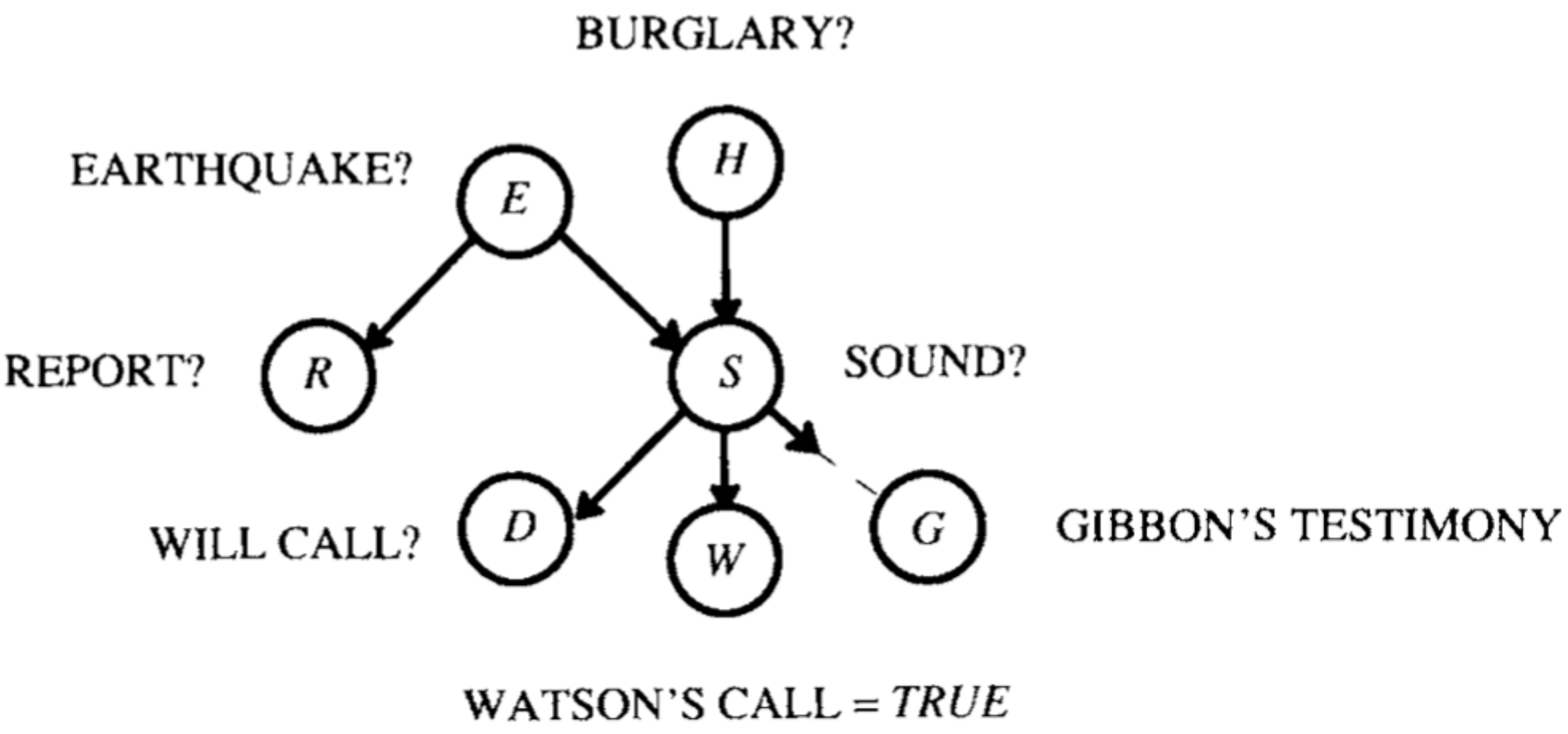
# Graphical Models and Variational Inference

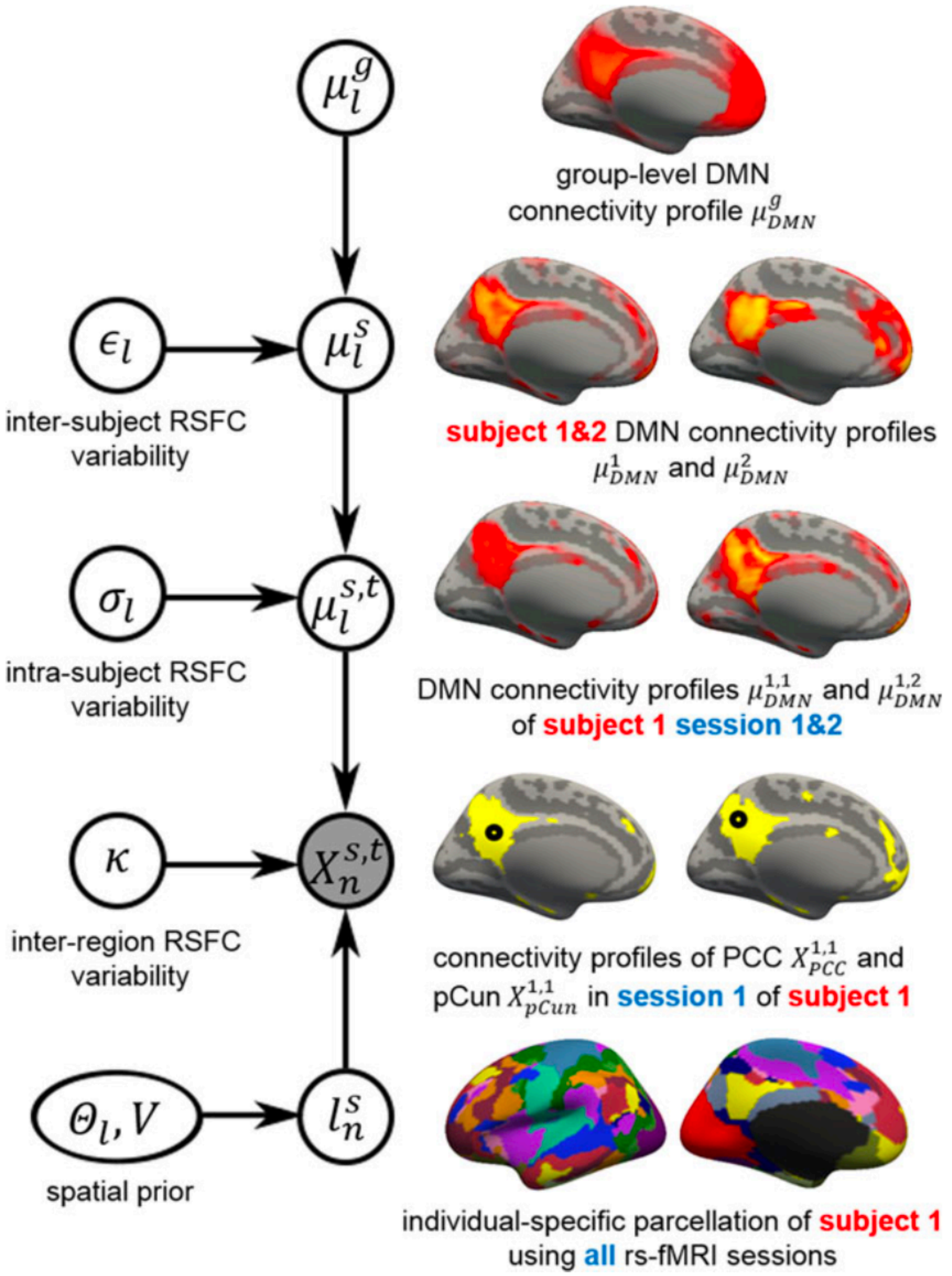
Demian Wassermann, Inria

Graphical Models: Discrete Inference and Learning

# Introduction to DAG and their relationship with Probability Functions (Pearl)

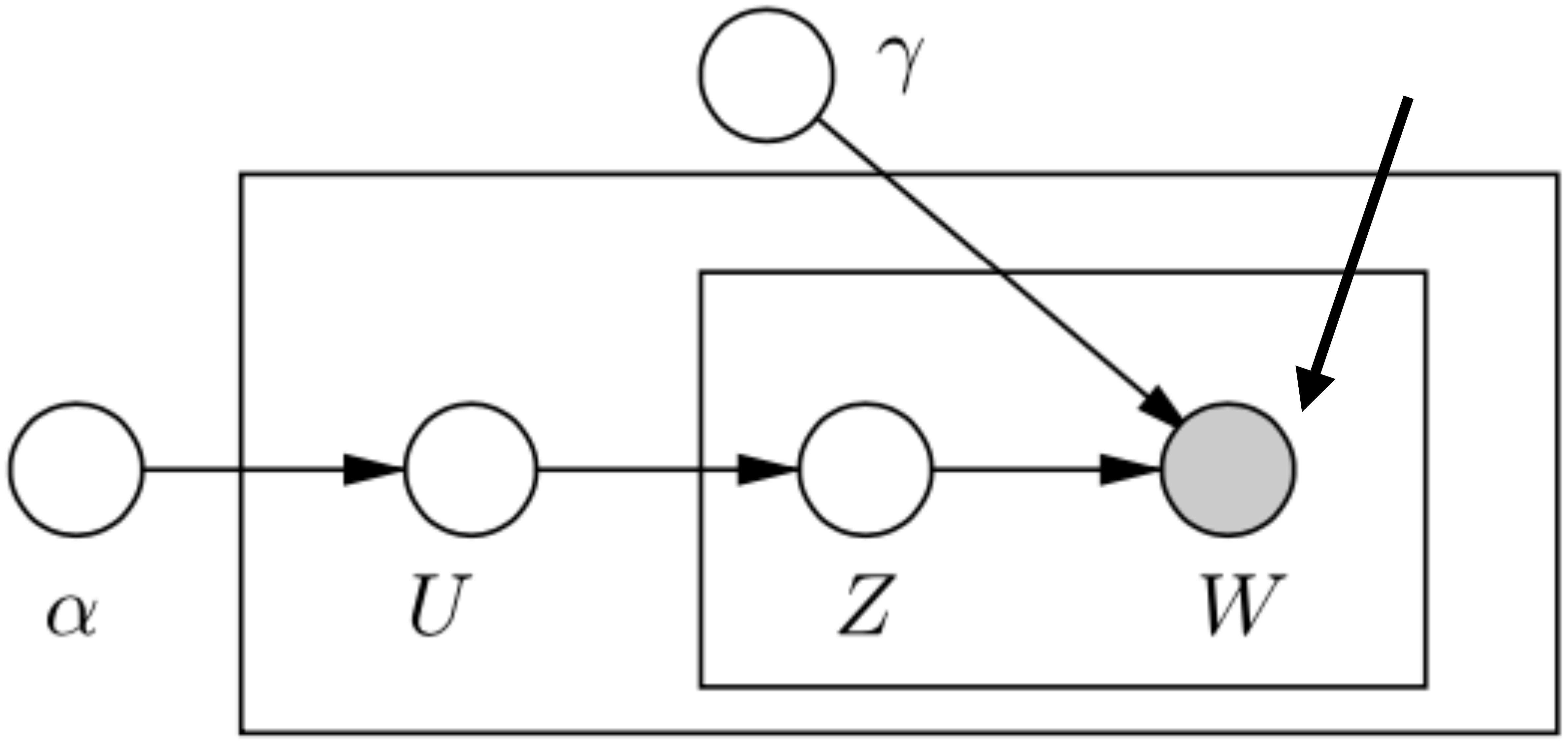



[Pearl 1987]




[Kong et al 2019]

# Introduction to DAG and their relationship with Probability Functions (Pearl)




U: is a Dirichlet or "clustering variable"
Z: is a "Topic"
W: is an *observed* "Word"
[Blei et al 2003]

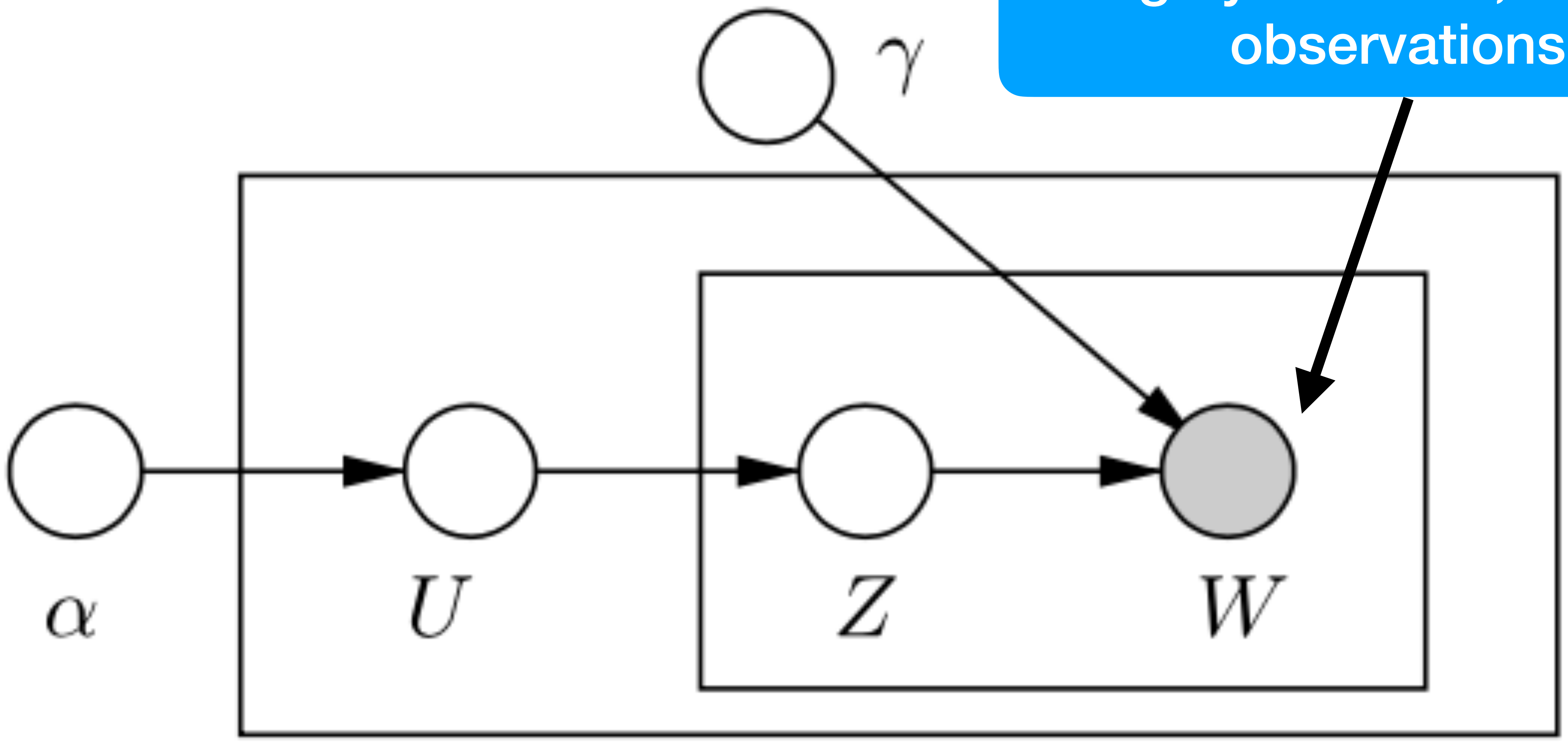
Each "box" or template represents a set of i.i.d. random variables with the same distribution

| "Arts" | "Budgets" | "Children" | "Education" |
|---|---|---|---|
| NEW | MILLION | CHILDREN | SCHOOL |
| FILM | TAX | WOMEN | STUDENTS |
| SHOW | PROGRAM | PEOPLE | SCHOOLS |
| MUSIC | BUDGET | CHILD | EDUCATION |
| MOVIE | BILLION | YEARS | TEACHERS |
| PLAY | FEDERAL | FAMILIES | HIGH |
| MUSICAL | YEAR | WORK | PUBLIC |
| BEST | SPENDING | PARENTS | TEACHER |
| ACTOR | NEW | SAYS | BENNETT |
| FIRST | STATE | FAMILY | MANIGAT |
| YORK | PLAN | WELFARE | NAMPHY |
| OPERA | MONEY | MEN | STATE |
| THEATER | PROGRAMS | PERCENT | PRESIDENT |
| ACTRESS | GOVERNMENT | CARE | ELEMENTARY |
| LOVE | CONGRESS | LIFE | HAITI |

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

# Introduction to DAG and their relationship with Probability Functions (Pearl)

The greyed nodes, represent observations




| "Arts" | "Budgets" | "Children" | "Education" |
|---|---|---|---|
| NEW | MILLION | CHILDREN | SCHOOL |
| FILM | TAX | WOMEN | STUDENTS |
| SHOW | PROGRAM | PEOPLE | SCHOOLS |
| MUSIC | BUDGET | CHILD | EDUCATION |
| MOVIE | BILLION | YEARS | TEACHERS |
| PLAY | FEDERAL | FAMILIES | HIGH |
| MUSICAL | YEAR | WORK | PUBLIC |
| BEST | SPENDING | PARENTS | TEACHER |
| ACTOR | NEW | SAYS | BENNETT |
| FIRST | STATE | FAMILY | MANIGAT |
| YORK | PLAN | WELFARE | NAMPHY |
| OPERA | MONEY | MEN | STATE |
| THEATER | PROGRAMS | PERCENT | PRESIDENT |
| ACTRESS | GOVERNMENT | CARE | ELEMENTARY |
| LOVE | CONGRESS | LIFE | HAITI |

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

U: is a Dirichlet or "clustering variable"

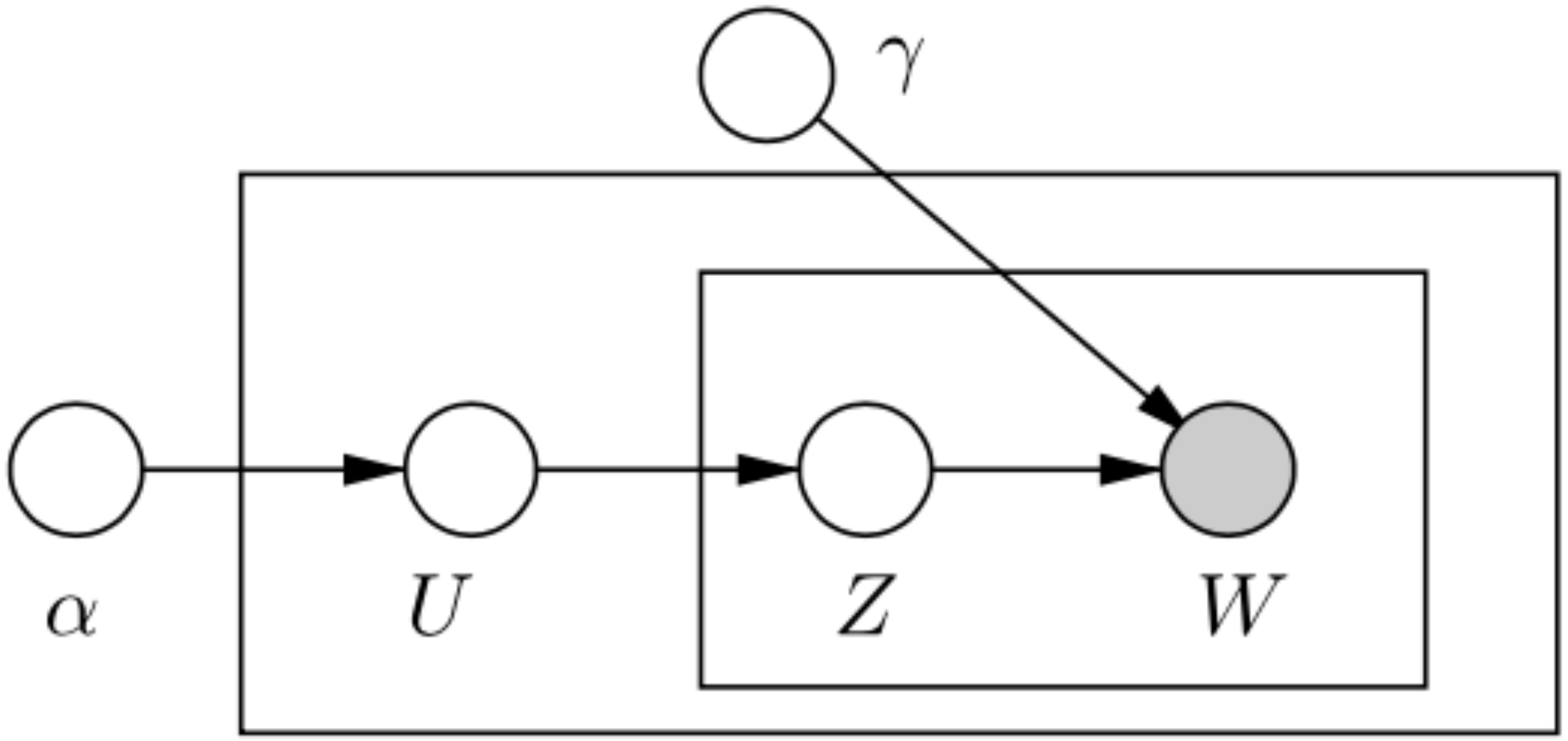
Z: is a "Topic"

W: is an *observed* "Word"

[Blei et al 2003]

Each "box" or template represents a set of i.i.d. random variables with the same distribution

# Introduction to DAG and their relationship with Probability Functions (Pearl)

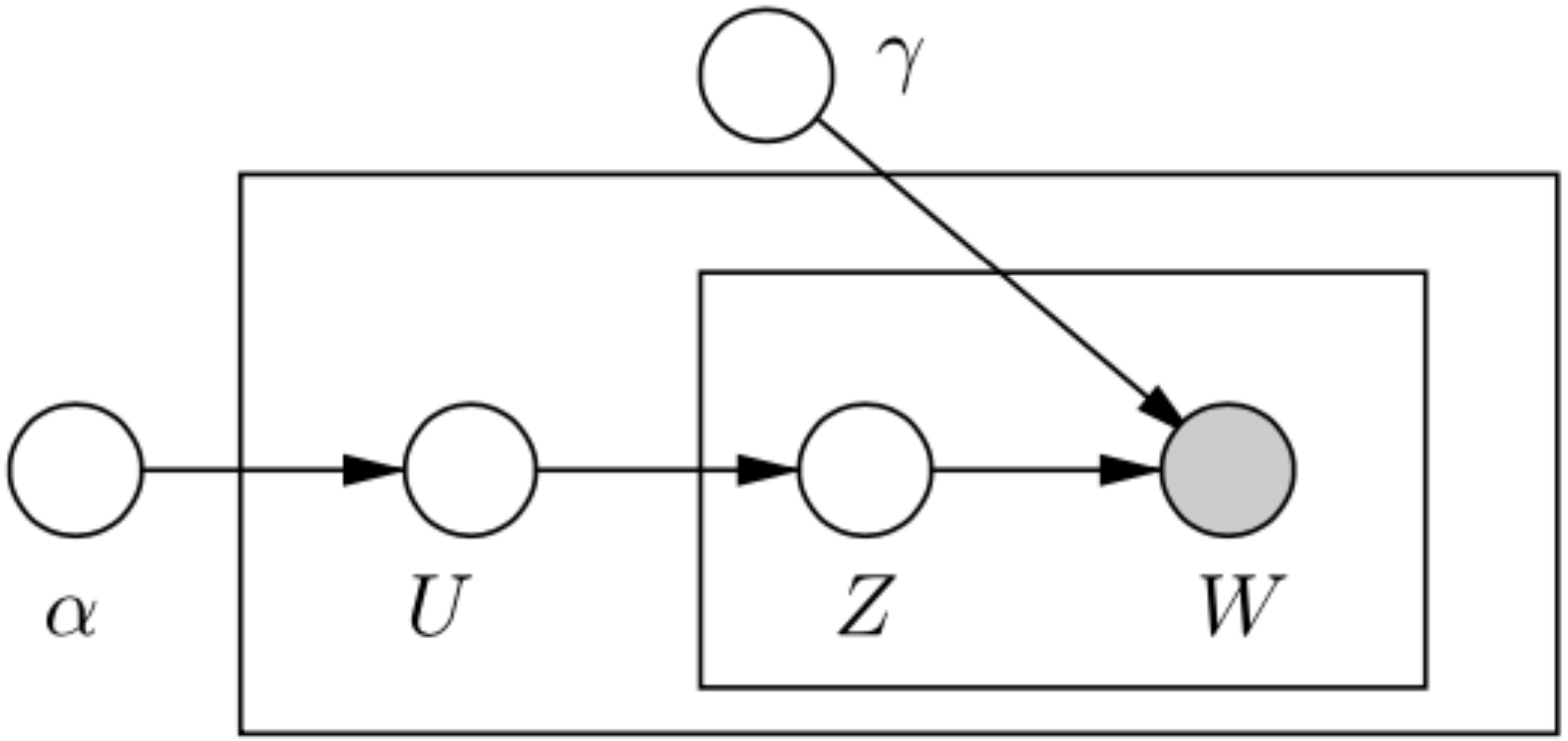



$$U_j \sim Dirichlet(\alpha), \alpha < 1$$
$$Z_{i,j} \sim Multinomial(U_j)$$
$$W_{i,j} \sim Multinomial\left(\gamma_{Z_{i,j}}\right)$$

Then, we are looking for the posterior $P(U, Z \mid W, \alpha, \gamma) = \dfrac{P(U, Z, W \mid \alpha, \gamma)}{P(W \mid \alpha, \gamma)}$

$$P(U, Z, W \mid \alpha, \gamma) = \Pi_j \int P(U_j \mid \alpha) \left( \Pi_i \sum_{Z_{i,j}} P(Z_{i,j} \mid U_j) P(W_{i,j} \mid Z_{i,j}, \gamma) \right) dU_j$$

# Introduction to DAG and their relationship with Probability Functions (Pearl)




$U_j \sim Dirichlet(\alpha), \alpha < 1$

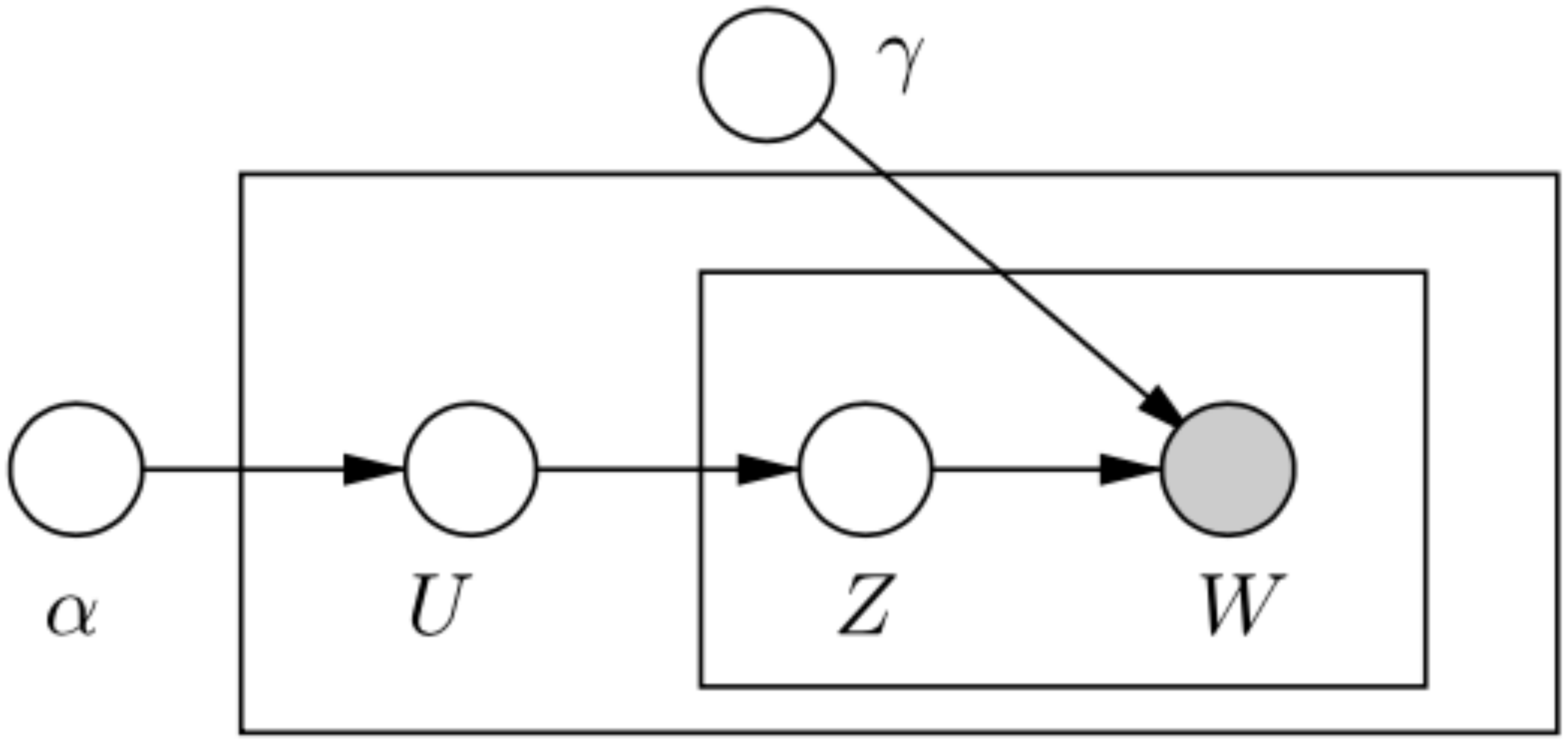
$Z_{i,j} \sim Multinomial(U_j)$

$W_{i,j} \sim Multinomial\left(\gamma_{Z_{i,j}}\right)$

Then, we are looking for the posterior $P(U, Z \mid W, \alpha, \gamma) = \dfrac{P(U, Z, W \mid \alpha, \gamma)}{P(W \mid \alpha, \gamma)}$

$$P(U, Z, W \mid \alpha, \gamma) = \Pi_j \int P(U_j \mid \alpha) \left( \Pi_i \sum_{Z_{i,j}} P(Z_{i,j} \mid U_j) P(W_{i,j} \mid Z_{i,j}, \gamma) \right) dU_j$$

# Introduction to DAG and their relationship with Probability Functions (Pearl)




$U_j \sim Dirichlet(\alpha), \alpha < 1$

$Z_{i,j} \sim Multinomial(U_j)$

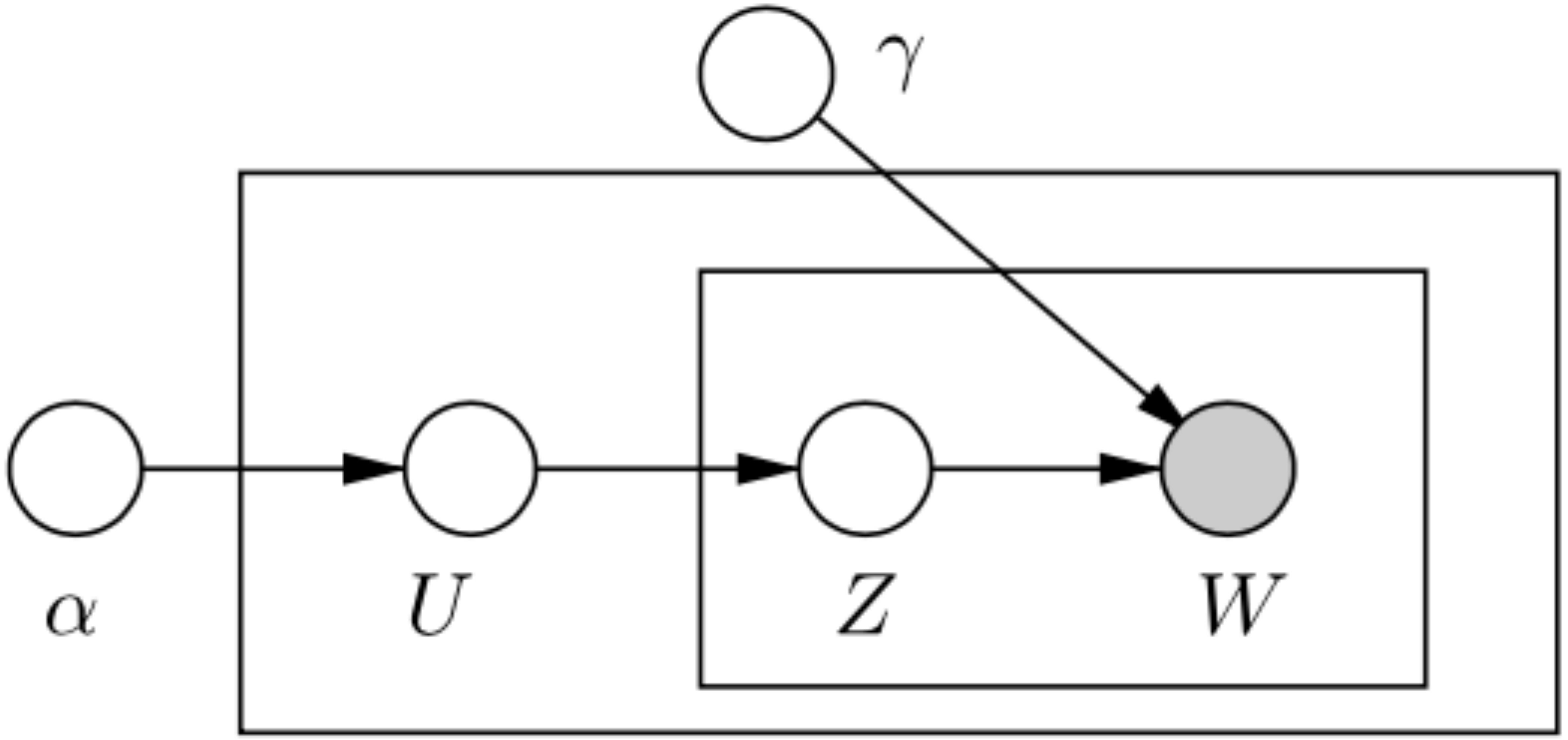
$W_{i,j} \sim Multinomial\left(\gamma_{Z_{i,j}}\right)$

Then, we are looking for the posterior $P(U, Z \mid W, \alpha, \gamma) = \dfrac{P(U, Z, W \mid \alpha, \gamma)}{P(W \mid \alpha, \gamma)}$

No analytical solution

$$P(U, Z, W \mid \alpha, \gamma) = \Pi_j \int P(U_j \mid \alpha) \left( \Pi_i \sum_{Z_{i,j}} P(Z_{i,j} \mid U_j) P(W_{i,j} \mid Z_{i,j}, \gamma) \right) dU_j$$

# Relationship between a Directed Graphical Model and its Probability Law (Pearl and Paz 1985)




| "Arts" | "Budgets" | "Children" | "Education" |
|---|---|---|---|
| NEW | MILLION | CHILDREN | SCHOOL |
| FILM | TAX | WOMEN | STUDENTS |
| SHOW | PROGRAM | PEOPLE | SCHOOLS |
| MUSIC | BUDGET | CHILD | EDUCATION |
| MOVIE | BILLION | YEARS | TEACHERS |
| PLAY | FEDERAL | FAMILIES | HIGH |
| MUSICAL | YEAR | WORK | PUBLIC |
| BEST | SPENDING | PARENTS | TEACHER |
| ACTOR | NEW | SAYS | BENNETT |
| FIRST | STATE | FAMILY | MANIGAT |
| YORK | PLAN | WELFARE | NAMPHY |
| OPERA | MONEY | MEN | STATE |
| THEATER | PROGRAMS | PERCENT | PRESIDENT |
| ACTRESS | GOVERNMENT | CARE | ELEMENTARY |
| LOVE | CONGRESS | LIFE | HAITI |

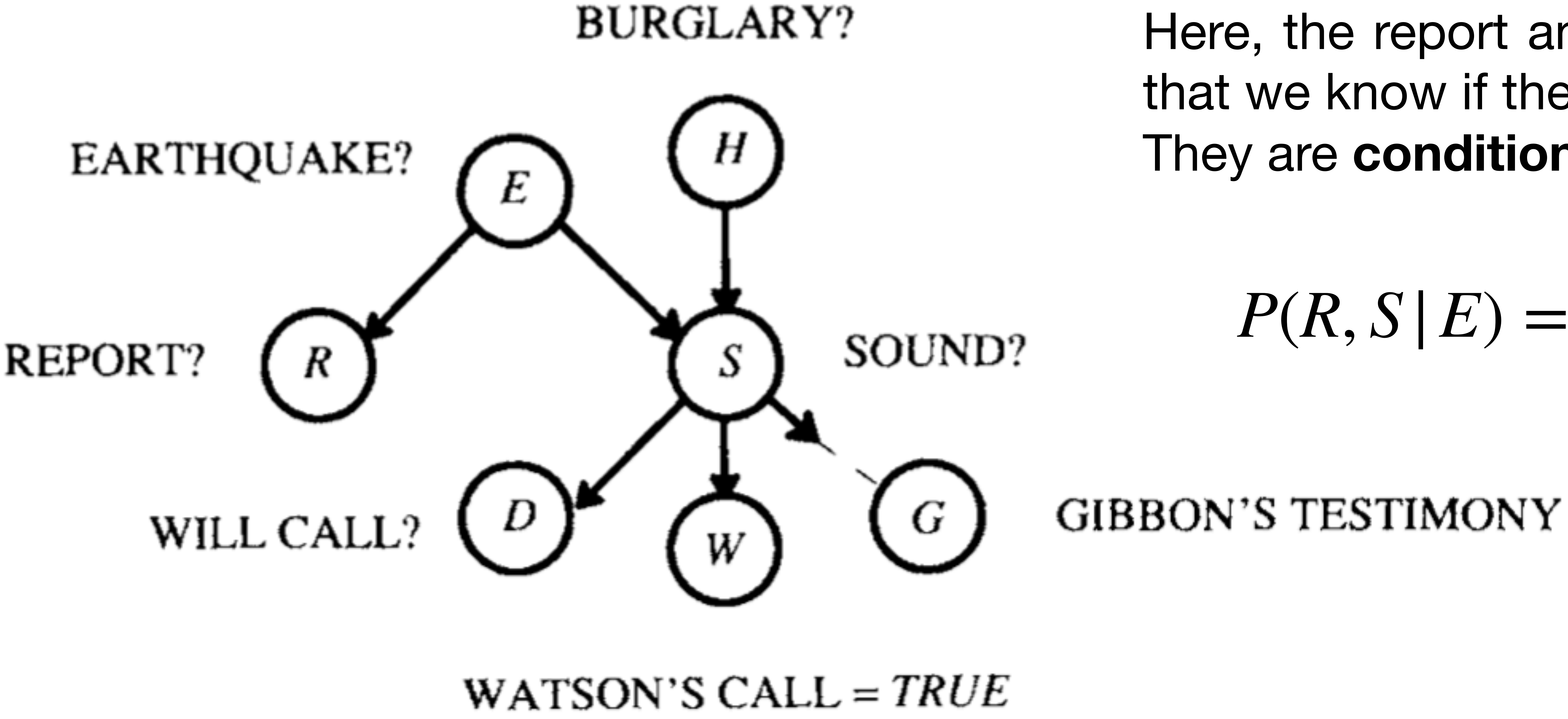
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

$$P(W_1, \ldots, W_I, Z_1, \ldots, Z_I, U_1 \ldots, U_J, \alpha, \gamma) = \Pi_j \Pi_i P(W_i \mid Z_i, \gamma) P(Z_i \mid U_j) P(U_j \mid \alpha)$$

In general, for a graphical model Graphical Model with vertices V and edges E

$$GM = (V, E), P(V) = \Pi_{v \in V} P(v \mid Pa(v)), Pa(v) = \{v' : v' \to v \in E\}$$

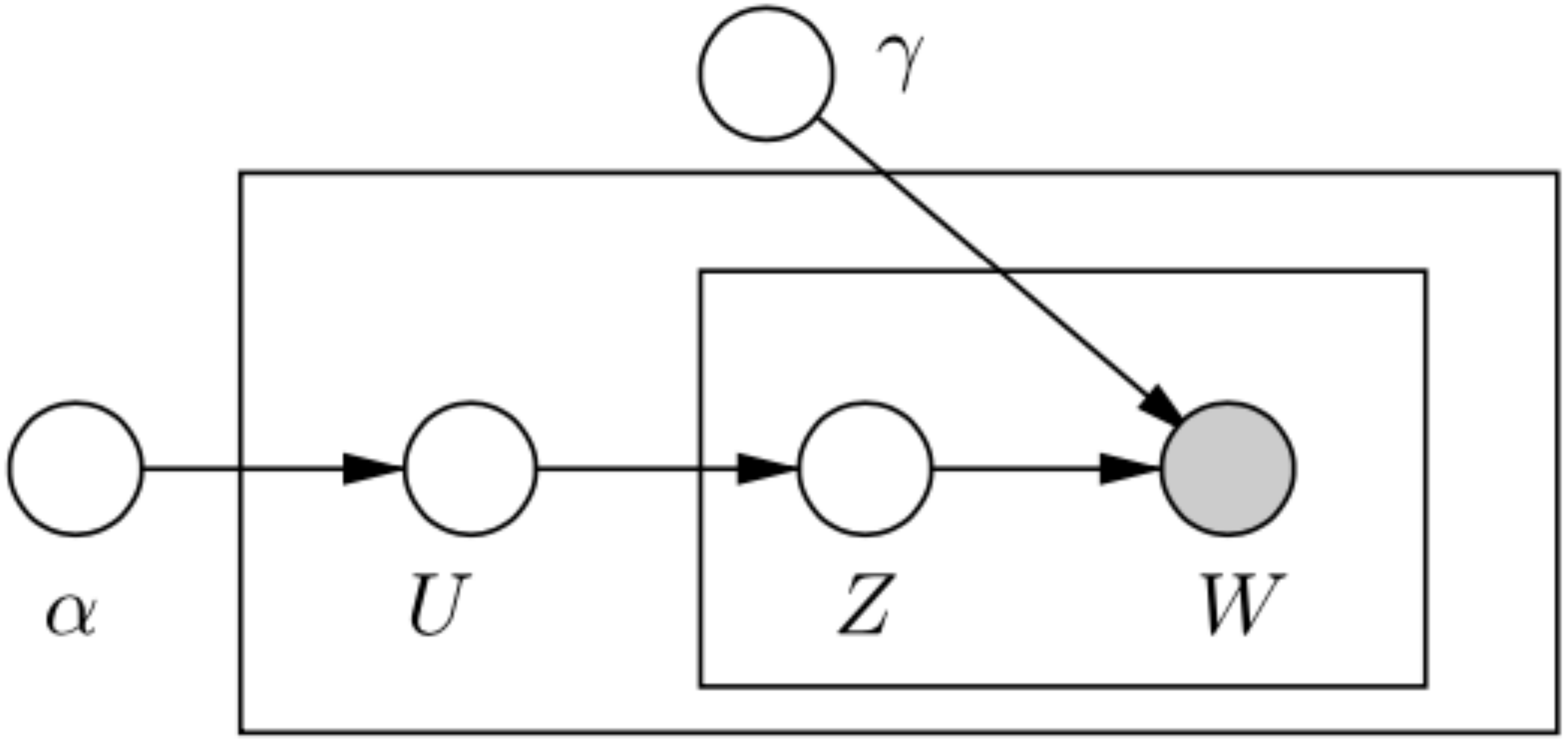
# Relationship between a Directed Graphical Model and its Probability Law (Pearl and Paz 1985)




Here, the report and the sound are independent, given that we know if there was an earthquake:
They are **conditionally** independent

$$P(R, S \mid E) = P(R \mid E)P(S \mid E) \text{ iif } I(R, S, E)$$

# Relationship between a Directed Graphical Model and its Probability Law (Pearl and Paz 1985)




| "Arts" | "Budgets" | "Children" | "Education" |
|---|---|---|---|
| NEW | MILLION | CHILDREN | SCHOOL |
| FILM | TAX | WOMEN | STUDENTS |
| SHOW | PROGRAM | PEOPLE | SCHOOLS |
| MUSIC | BUDGET | CHILD | EDUCATION |
| MOVIE | BILLION | YEARS | TEACHERS |
| PLAY | FEDERAL | FAMILIES | HIGH |
| MUSICAL | YEAR | WORK | PUBLIC |
| BEST | SPENDING | PARENTS | TEACHER |
| ACTOR | NEW | SAYS | BENNETT |
| FIRST | STATE | FAMILY | MANIGAT |
| YORK | PLAN | WELFARE | NAMPHY |
| OPERA | MONEY | MEN | STATE |
| THEATER | PROGRAMS | PERCENT | PRESIDENT |
| ACTRESS | GOVERNMENT | CARE | ELEMENTARY |
| LOVE | CONGRESS | LIFE | HAITI |

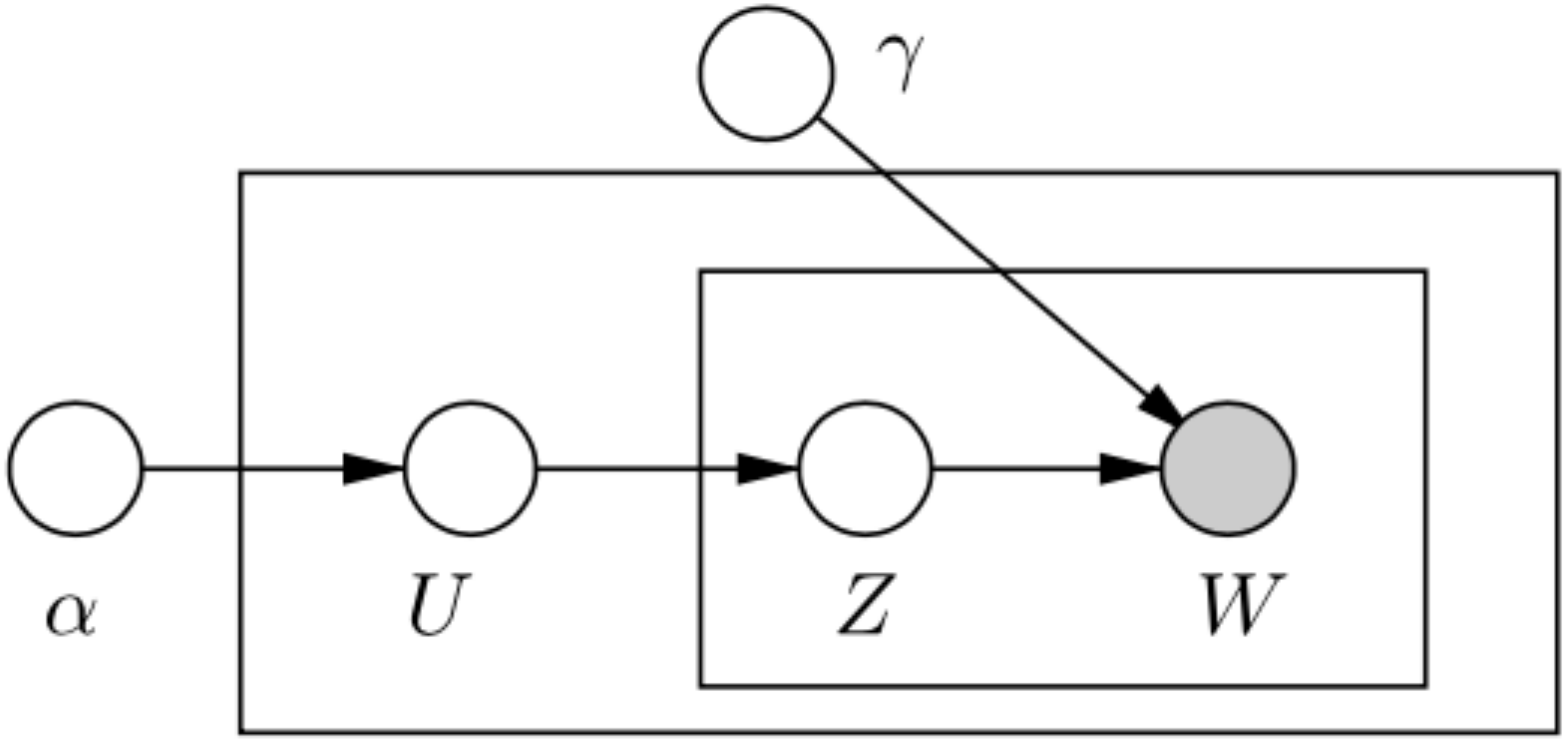
The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

$$P(W_1, \dots, W_I, Z_1, \dots, Z_I, U_1 \dots, U_J, \alpha, \gamma) = \Pi_j \Pi_i P(W_i \mid Z_i, \gamma) P(Z_i \mid U_j) P(U_j \mid \alpha)$$

However, our usual problem is: given observed variables O and latent variables L, to compute the posterior $P(L \mid O)$

$$P(L \mid O) = \frac{\Pi_{v \in V} P(v \mid Pa(v))}{\Pi_o P(o \mid Pa(o))}, GM = (V = L \cup O, E), \;\; \nexists l \in L : o \to l \in E$$

# Relationship between a Directed Graphical Model and its Probability Law (Pearl and Paz 1985)




| "Arts" | "Budgets" | "Children" | "Education" |
|---|---|---|---|
| NEW | MILLION | CHILDREN | SCHOOL |
| FILM | TAX | WOMEN | STUDENTS |
| SHOW | PROGRAM | PEOPLE | SCHOOLS |
| MUSIC | BUDGET | CHILD | EDUCATION |
| MOVIE | BILLION | YEARS | TEACHERS |
| PLAY | FEDERAL | FAMILIES | HIGH |
| MUSICAL | YEAR | WORK | PUBLIC |
| BEST | SPENDING | PARENTS | TEACHER |
| ACTOR | NEW | SAYS | BENNETT |
| FIRST | STATE | FAMILY | MANIGAT |
| YORK | PLAN | WELFARE | NAMPHY |
| OPERA | MONEY | MEN | STATE |
| THEATER | PROGRAMS | PERCENT | PRESIDENT |
| ACTRESS | GOVERNMENT | CARE | ELEMENTARY |
| LOVE | CONGRESS | LIFE | HAITI |

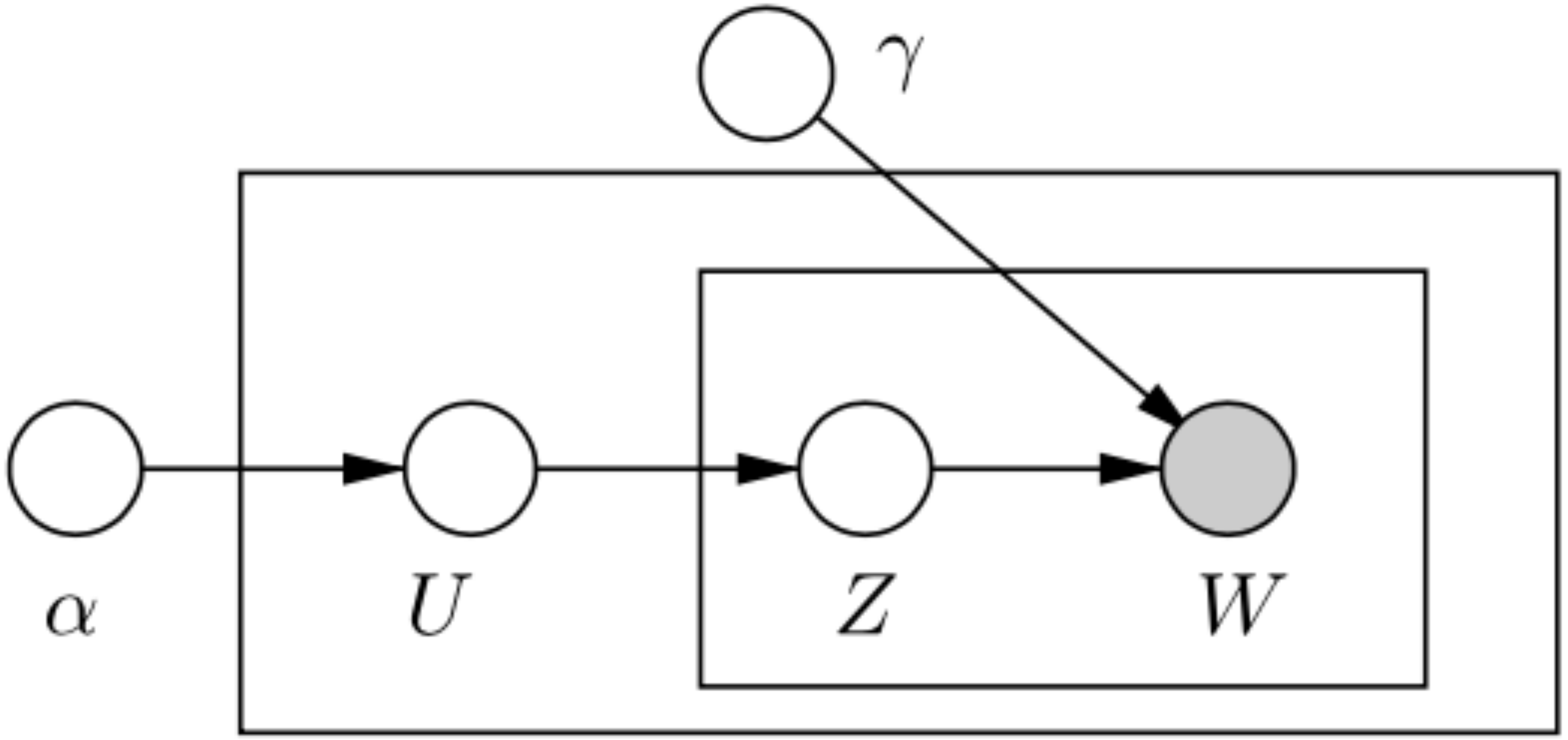
The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

$$P(L \mid O) = \frac{\Pi_{v \in V} P(v \mid Pa(v))}{\Pi_o P(o \mid Pa(o))}, GM = (V = L \cup O, E), \;\; \nexists l \in L : o \to l \in E$$

In the case of continuous variables this is

$$P(L \mid O) = \frac{P(L, O)}{\int P(L, O) dO}$$

# Relationship between a Directed Graphical Model and its Probability Law (Pearl and Paz 1985)




| "Arts" | "Budgets" | "Children" | "Education" |
|---|---|---|---|
| NEW | MILLION | CHILDREN | SCHOOL |
| FILM | TAX | WOMEN | STUDENTS |
| SHOW | PROGRAM | PEOPLE | SCHOOLS |
| MUSIC | BUDGET | CHILD | EDUCATION |
| MOVIE | BILLION | YEARS | TEACHERS |
| PLAY | FEDERAL | FAMILIES | HIGH |
| MUSICAL | YEAR | WORK | PUBLIC |
| BEST | SPENDING | PARENTS | TEACHER |
| ACTOR | NEW | SAYS | BENNETT |
| FIRST | STATE | FAMILY | MANIGAT |
| YORK | PLAN | WELFARE | NAMPHY |
| OPERA | MONEY | MEN | STATE |
| THEATER | PROGRAMS | PERCENT | PRESIDENT |
| ACTRESS | GOVERNMENT | CARE | ELEMENTARY |
| LOVE | CONGRESS | LIFE | HAITI |

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

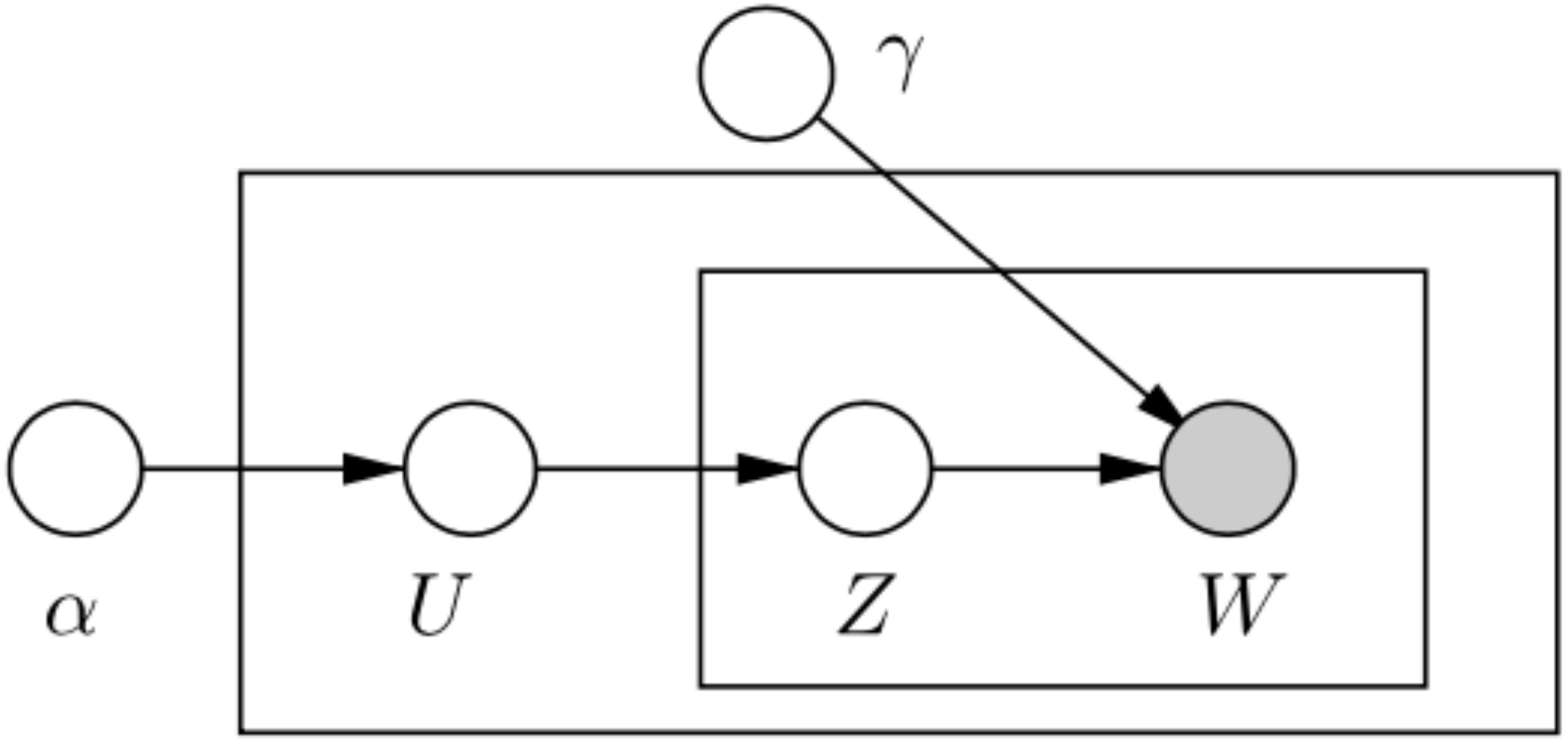
$$P(L\,|\,O) = \frac{\Pi_{v\in V} P(v\,|\,Pa(v))}{\Pi_o P(o\,|\,Pa(o))}, GM = (V = L \cup O, E),\ \nexists l \in L : o \to l \in E$$

In the case of continuous variables this is

**No analytical solution, for the general case**

$$P(L\,|\,O) = \frac{P(L,O)}{\int P(L,O)dO}$$

# Relationship between a Directed Graphical Model and its Probability Law (Pearl and Paz 1985)




| "Arts" | "Budgets" | "Children" | "Education" |
|---|---|---|---|
| NEW | MILLION | CHILDREN | SCHOOL |
| FILM | TAX | WOMEN | STUDENTS |
| SHOW | PROGRAM | PEOPLE | SCHOOLS |
| MUSIC | BUDGET | CHILD | EDUCATION |
| MOVIE | BILLION | YEARS | TEACHERS |
| PLAY | FEDERAL | FAMILIES | HIGH |
| MUSICAL | YEAR | WORK | PUBLIC |
| BEST | SPENDING | PARENTS | TEACHER |
| ACTOR | NEW | SAYS | BENNETT |
| FIRST | STATE | FAMILY | MANIGAT |
| YORK | PLAN | WELFARE | NAMPHY |
| OPERA | MONEY | MEN | STATE |
| THEATER | PROGRAMS | PERCENT | PRESIDENT |
| ACTRESS | GOVERNMENT | CARE | ELEMENTARY |
| LOVE | CONGRESS | LIFE | HAITI |

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

$$P(L\,|\,O) = \frac{\Pi_{v\in V} P(v\,|\,Pa(v))}{\Pi_o P(o\,|\,Pa(o))}, GM = (V = L \cup O, E),\ \nexists l \in L : o \to l \in E$$

Can we approximate $P(L\,|\,O)$?

$$\color{blue}{Q(L)} \simeq P(L\,|\,O) = \frac{P(L, O)}{\int P(L, O) dO}$$

# Approximations to Density Laws

Can we approximate $P(L \mid O)$? $\quad Q(L) \simeq P(L \mid O) = \dfrac{P(L,O)}{\int P(L,O)dO}$

- First try: MacLaurin $Q(L) = \sum P(L = l \mid O) + P'(L = l \mid O)(l - L) + \dots$

  problem: how to guarantee that $Q(L)$ is a probability law?

- Second try: cumulant approximations (changing the random $L$ by $X$)

$$\phi(t) = \log \mathbb{E}_{X \sim Q(X)}[\exp(tX)] = \sum_n \kappa_n \frac{t^n}{n!} = \kappa_1 t + \kappa_2 \frac{t^2}{2!} + \dots = \mu t + \sigma^2 \frac{t^2}{2!} + \dots$$

# Approximations to Density Laws

Can we approximate $P(L \mid O)$?

$$Q(L) \simeq P(L \mid O) = \frac{P(L, O)}{\int P(L, O) dO}$$

- First try: MacLaurin $Q(L) = \sum P(L = l \mid O) + P'(L = l \mid O)(l - L) + \ldots$

  problem: how to guarantee that $Q(L)$ is a probability law?

- Second try: cumulant approximations (changing the random $L$ by $X$)

  $$\phi(t) = \log \mathbb{E}_{X \sim Q(X)}[\exp(tX)] = \sum_n \kappa_n \frac{t^n}{n!} = \kappa_1 t + \kappa_2 \frac{t^2}{2!} + \ldots = \mu t + \sigma^2 \frac{t^2}{2!} + \ldots$$

- However, a probability law has either up to two moments, or an infinite number (Cramèr 1938)

# Approximations to Density Laws

Can we approximate $P(L \mid O)$?

$$Q(L) \simeq P(L \mid O) = \frac{P(L, O)}{\int P(L, O) dO}$$

- Other options: Edgesworth, approximations which come from this identity

$$\phi(t) = \log \mathbb{E}_X[\exp(itX)] = \sum_n \kappa_n \frac{(it)^n}{n!},$$

$$\psi(t) = \log \mathbb{E}_X[\exp(itX)] = \sum_n \gamma_n \frac{(it)^n}{n!}$$

$$\hat{\phi}(t) = \sum_n (\kappa_n - \gamma_n) \frac{(it)^n}{n!} + \log \psi(t)$$

however, they are not guaranteed to be probability laws for finite samples.

# Approximations to Density Laws

Can we approximate $P(L \,|\, O)$?

$$\color{blue}{Q(L)} \simeq P(L \,|\, O) = \frac{P(L, O)}{\int P(L, O) dO}$$

- So? What do we do?
  - We choose an approximate distribution $Q_\theta(X)$ —replacing $L$ by $X$ and $O$ by $Z$ for notation— from a given family, with parameters $\theta$. Then
    $$Q* = Q_{\theta*} : \theta* = \arg\min_\theta D(Q_\theta(X), P(X \,|\, Z))$$
    so we need to define the right similarity measurement $D$ to compare distributions. And in standard Variational Inference (VI), $Z$ is notation for $O$

# Approximations to Dens… …aws

Can we approximate $P(L|O)$?

$\color{blue}{Q(L)} \simeq P(L|O)$ …

This is what we call Variational Inference

- So? What do we do?
  - We choose an approximate distribution $Q_\theta(X)$ —replacing $L$ by $X$ an… by $Z$ for notation— from a given family, with parameters $\theta$. Then

$$Q* = Q_{\theta*} : \theta* = \arg\min_\theta D(Q_\theta(X), P(X|Z))$$

so we need to define the right similarity measurement $D$ to compare distributions. And in standard Variational Inference (VI), $Z$ is notation for $O$

# So Which $D$ and $Q$ Should We Choose?

$$Q* = Q_{\theta*} : \theta* = \arg\min_{\theta} D(Q_\theta(X), P(X|Z))$$

$X$ the latent variables and $Z$ the observations

Let's start with "analytical" ideas:

- $D(Q_\theta(X), P(X|Z)) = \int (Q_\theta(x) - P(x|Z))^2 dx$
  - What does it mean for two distributions to be close in the $L_2$ sense?
  - How easy is to obtain bounds and closed form solutions?

- $Q_\theta(X) : X \sim \mathcal{N}(\mu, \Sigma), \theta = (\mu, \Sigma)$ : This is called the Laplace approximation
  - Even simpler $\Sigma = \sigma^2$Id, which boils down to $Q_\mu(X) = \Pi_i Q_{\mu_i}(X_i)$

# So Which $D$ and $Q$ Should We Choose?

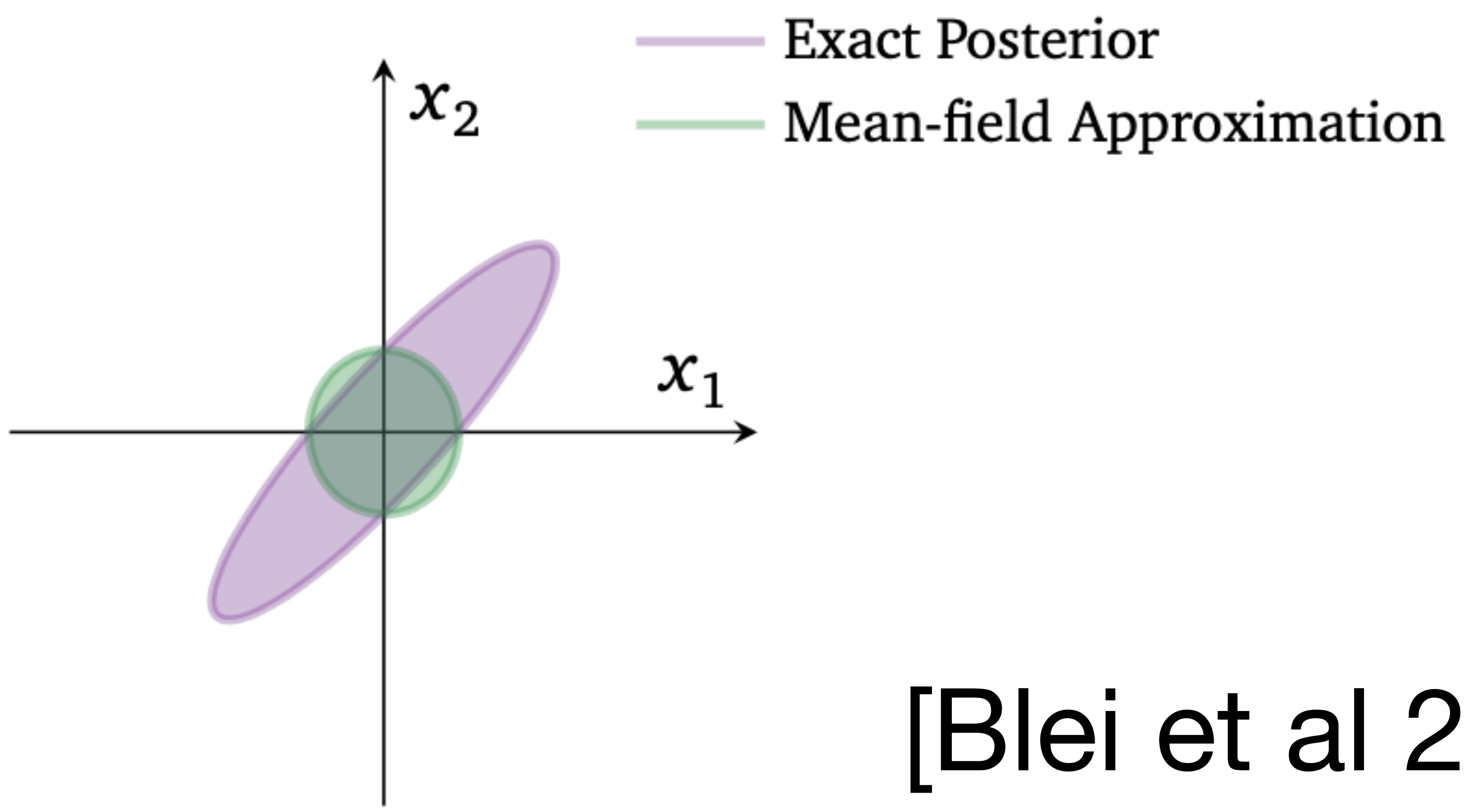
$$Q^* = Q_{\theta^*} : \theta^* = \arg\min_{\theta} D(Q_\theta(X), P(X|Z))$$

$X$ the latent variables and $Z$ the observations

More Information theoretic

- $$D_{KL}(Q_\theta(X), P(X|Z)) = \mathbb{E}_{X \sim Q_\theta}\left[-\log \frac{P(X|Z)}{Q_\theta(X)}\right] = -\int dQ_\theta(x) \log \frac{P(x|Z)}{Q_\theta(x)}$$
  - The Kullback-Leibler divergence is based on information theory
  - Known formulations for common cases

- Mean field $Q_{\theta=\mu}(X) = \Pi_i Q_{\mu_i}(X_i)$




[Blei et al 2017]

# A Case for Mean Field KL-based VI

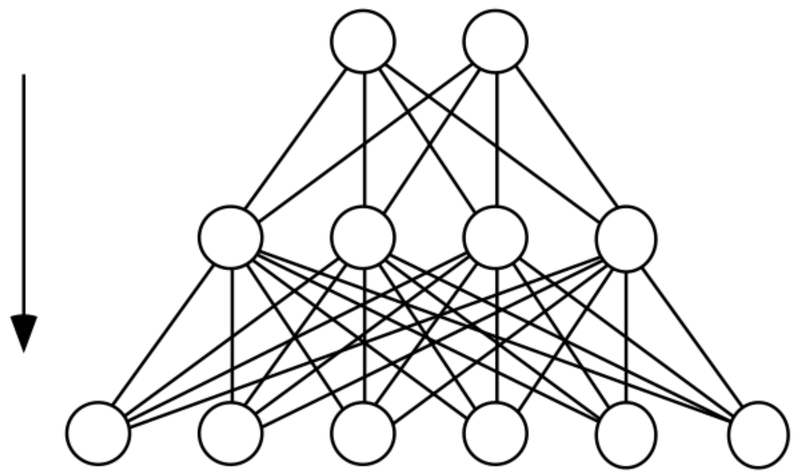
Journal of Artificial Intelligence Research 4 (1996) 61—76 Submitted 11/95; published 3/96

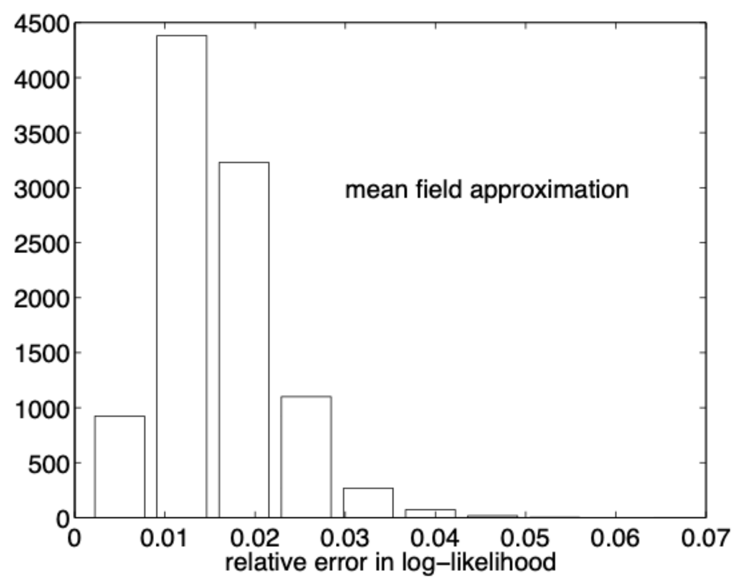
## Mean Field Theory for Sigmoid Belief Networks

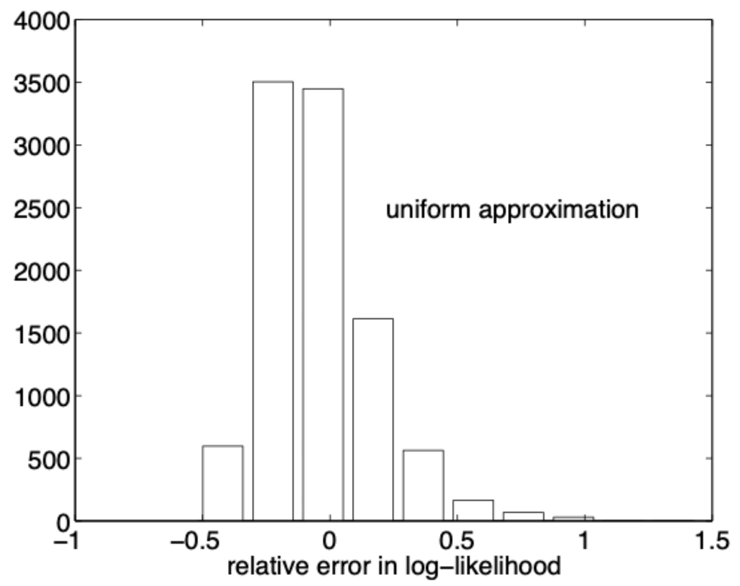
**Lawrence K. Saul** LKSAUL@PSYCHE.MIT.EDU
**Tommi Jaakkola** TOMMI@PSYCHE.MIT.EDU
**Michael I. Jordan** JORDAN@PSYCHE.MIT.EDU
*Center for Biological and Computational Learning*
*Massachusetts Institute of Technology*
*79 Amherst Street, E10-243*
*Cambridge, MA 02139*

**Abstract**

We develop a mean field theory for sigmoid belief networks based on ideas from statistical mechanics. Our mean field theory provides a tractable approximation to the true probability distribution in these networks; it also yields a lower bound on the likelihood of evidence. We demonstrate the utility of this framework on a benchmark problem in statistical pattern recognition—the classification of handwritten digits.





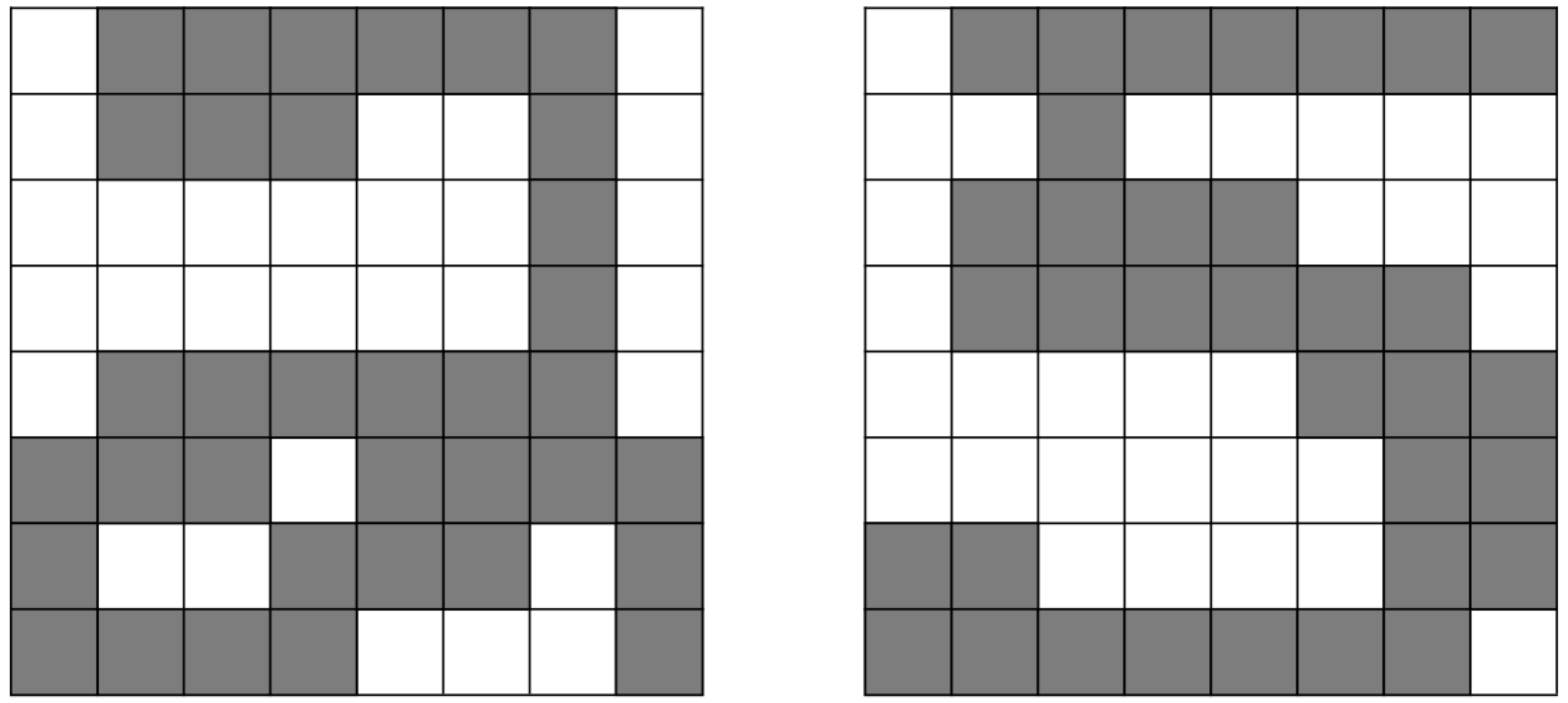



Figure 7: Binary images of handwritten digits: two and five.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 388 | 2 | 2 | 0 | 1 | 3 | 0 | 0 | 4 | 0 |
| 1 | 0 | 393 | 0 | 0 | 0 | 1 | 0 | 0 | 6 | 0 |
| 2 | 1 | 2 | 376 | 1 | 3 | 0 | 4 | 0 | 13 | 0 |
| 3 | 0 | 2 | 4 | 373 | 0 | 12 | 0 | 0 | 6 | 3 |
| 4 | 0 | 0 | 2 | 0 | 383 | 0 | 1 | 2 | 2 | 10 |
| 5 | 0 | 2 | 1 | 13 | 0 | 377 | 2 | 0 | 4 | 1 |
| 6 | 1 | 4 | 2 | 0 | 1 | 6 | 386 | 0 | 0 | 0 |
| 7 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 388 | 3 | 8 |
| 8 | 1 | 9 | 1 | 7 | 0 | 7 | 1 | 1 | 369 | 4 |
| 9 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 8 | 5 | 383 |

# So Which $D$ and $Q$ Should We Choose?

$$Q^* = Q_{\theta^*} : \theta^* = \arg\min_{\theta} D(Q_\theta(X), P(X|Z))$$

$X$ the latent variables and $Z$ the observations

A second order information-theoretic model

- $D_{KL}(Q_\theta(X), P(X|Z)) = \mathbb{E}_{X\sim Q_\theta}\left[-\log \frac{P(X|Z)}{Q_\theta(X)}\right] = -\int dQ_\theta(x)\log \frac{P(x|Z)}{Q_\theta(x)}$
- $Q_\theta(X) : X \sim \mathcal{N}(\mu, \Sigma), \theta = (\mu, \Sigma)$ : This is called the Laplace approximation

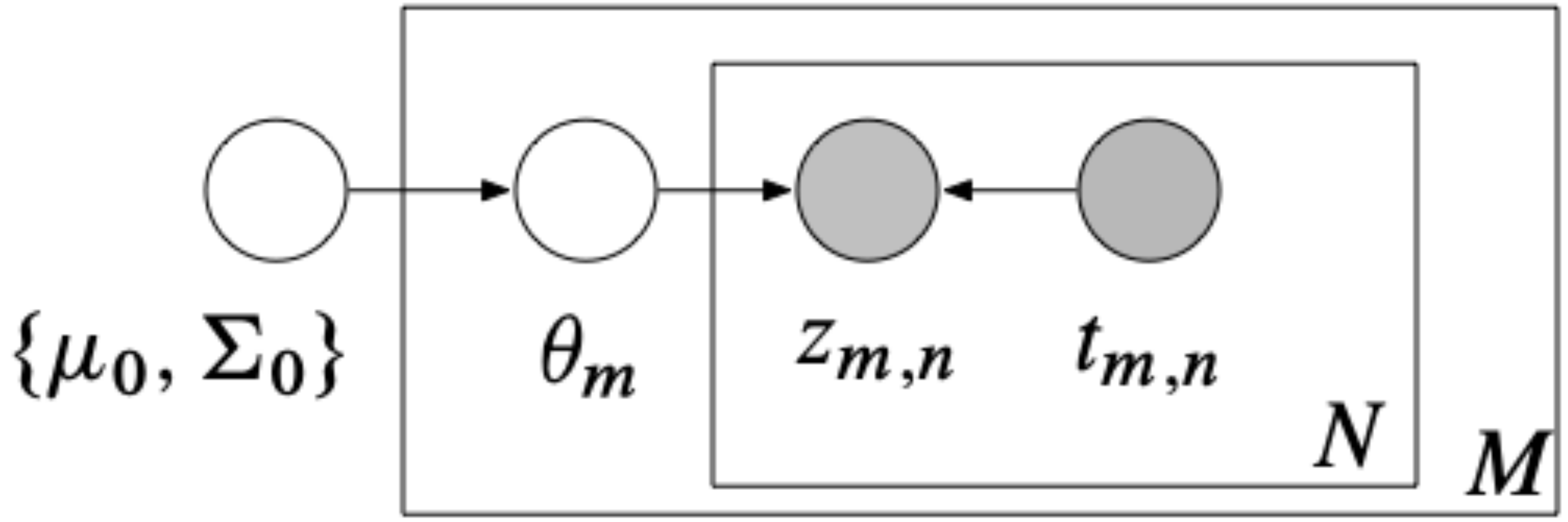
# But Laplace is Better

Journal of Machine Learning Research (2013) Submitted 00/00; Published 00/00

## Variational Inference in Nonconjugate Models

**Chong Wang** CHONGW@CS.CMU.EDU
*Machine Learning Department*
*Carnegie Mellon University*
*Pittsburgh, PA, 15213, USA*

**David M. Blei** BLEI@CS.PRINCETON.EDU
*Department of Computer Science*
*Princeton University*
*Princeton, NJ, 08540, USA*


1. Draw coefficients $\theta \sim \mathcal{N}(\mu_0, \Sigma_0)$.
2. For each data point $n$ and its covariates $t_n$, draw its class label from

$$z_n \,|\, \theta, t_n \sim \text{Bernoulli}\left(\sigma(\theta^\top t_n)^{z_{n,1}} \sigma(-\theta^\top t_n)^{z_{n,2}}\right),$$

| | *Yeast* | | *Scene* | |
|---|---|---|---|---|
| | Accuracy | Log Likelihood | Accuracy | Log Likelihood |
| Jaakkola and Jordan (1996) | 79.7% | -0.678 | 87.4% | -0.670 |
| Laplace inference | **80.1%** | **-0.449** | **89.4%** | **-0.259** |

# So Which $D$ and $Q$ Should We Choose?

$$Q^* = Q_{\theta^*} : \theta^* = \arg\min_{\theta} D(Q_\theta(X), P(X|Z))$$

$X$ the latent variables and $Z$ the observations

A second order information-theoretic model

- $D_{KL}(Q_\theta(X), P(X|Z)) = \mathbb{E}_{X \sim Q_\theta}\left[-\log \frac{P(X|Z)}{Q_\theta(X)}\right] = -\int dQ_\theta(x) \log \frac{P(x|Z)}{Q_\theta(x)}$
- $Q_\theta(X) : X \sim \mathcal{N}(\mu, \Sigma), \theta = (\mu, \Sigma)$ : This is called the Laplace approximation

# So Which $D$ Should We Choose? Finding Bounds

$$D_{KL}(Q_\theta(X), P(X|Z)) = \mathbb{E}_{X\sim Q_\theta}\left[-\log \frac{P(X|Z)}{Q_\theta(X)}\right] = -\int dQ_\theta(x) \log \frac{P(x|Z)}{Q_\theta(x)}$$

But our graphical model is more adapted to sample from $P(X,Z)$ than from $P(X|Z)$.

Then, can we find a way to efficiently minimise $D_{KL}\left(Q_\theta(X), \frac{P(X,Z)}{P(Z)}\right)$

when, in general, we don't know the probability of "evidence" $P(Z)$?

Let's see in the next slide….

# So Which $D$ Should We Choose? Finding Bounds

$$D_{KL}(Q_\theta(X), P(X|Z)) = \mathbb{E}_{X \sim Q_\theta}\left[-\log \frac{P(X|Z)}{Q_\theta(X)}\right] = -\int dQ_\theta(x) \log \frac{P(x|Z)}{Q_\theta(x)}$$

And we know that

$$\log P(Z) = \log \int dx P(x, Z) = \log \int \frac{dQ_\theta(x) P(x, Z)}{Q_\theta(x)} = \log \mathbb{E}_{X \sim Q_\theta}\left[\frac{P(X, Z)}{Q_\theta(X)}\right]$$

with $Z$ being the observed data ($O$ before) and $X$ our latent variables ($L$)

then, $P(Z) = \log \mathbb{E}_{X \sim Q_\theta}\left[\frac{P(X, Z)}{Q_\theta(X)}\right] \geq E_{X \sim Q_\theta}\left[\frac{P(X, Z)}{Q_\theta(X)}\right] \triangleq \mathscr{L}(\theta)$

$$\min_\theta D_{KL}(Q_\theta(X), P(X|Z)) = \log P(Z) - \max_\theta \mathscr{L}(\theta)$$

Hence, it is enough to maximise the Evidence Lower Bound (ELBO): $\mathscr{L}(\theta)$

# So Which $D$ and $Q$ Should We Choose?

$$Q^* = Q_{\theta^*} : \theta^* = \arg\min_{\theta} D(Q_\theta(X), P(X \mid Z))$$

$X$ the latent variables and $Z$ the observations

A simplified second order information-theoretic model

- $\theta = \arg\max_{\theta} \mathscr{L}(\theta) = \mathbb{E}_{X \sim Q_\theta}\left[\log \frac{P(X, Z)}{Q_\theta(X)}\right]$
- $Q_\theta(X) : X \sim \mathcal{N}(\mu, \Sigma), \theta = (\mu, \Sigma)$ : This is called the Laplace approximation

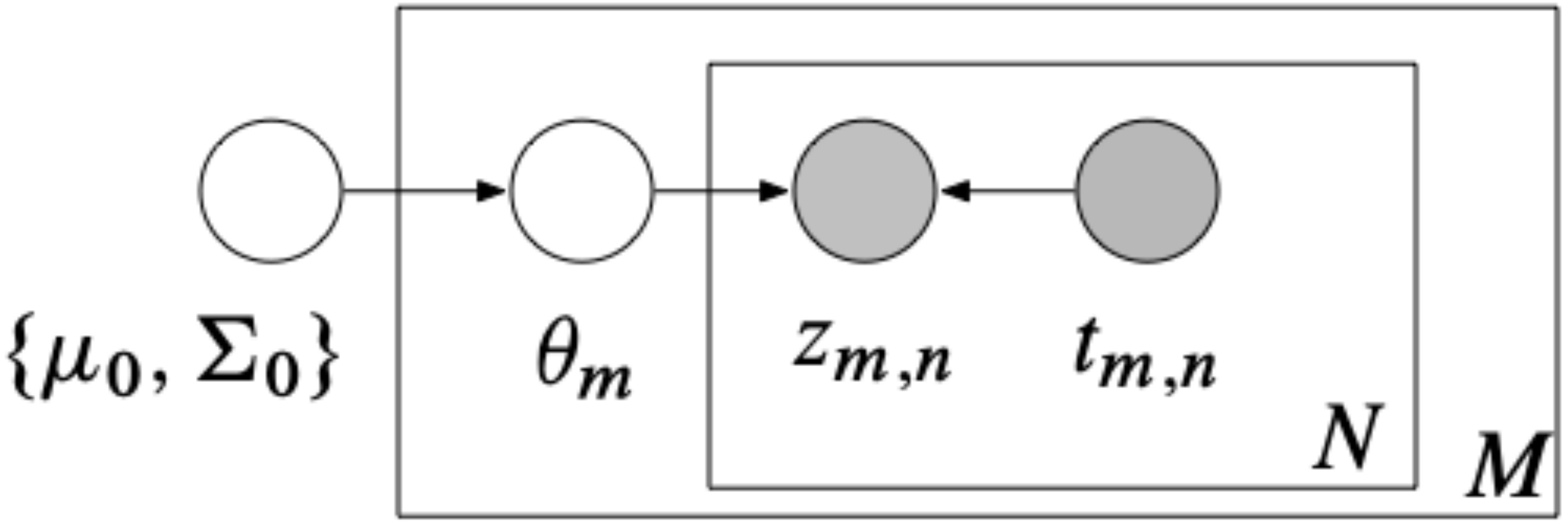
# But Laplace is Better (they use ELBO)

Journal of Machine Learning Research (2013) Submitted 00/00; Published 00/00

## Variational Inference in Nonconjugate Models

**Chong Wang** CHONGW@CS.CMU.EDU
*Machine Learning Department*
*Carnegie Mellon University*
*Pittsburgh, PA, 15213, USA*

**David M. Blei** BLEI@CS.PRINCETON.EDU
*Department of Computer Science*
*Princeton University*
*Princeton, NJ, 08540, USA*


1. Draw coefficients $\theta \sim \mathcal{N}(\mu_0, \Sigma_0)$.
2. For each data point $n$ and its covariates $t_n$, draw its class label from

$$z_n \mid \theta, t_n \sim \text{Bernoulli}\left(\sigma(\theta^\top t_n)^{z_{n,1}} \sigma(-\theta^\top t_n)^{z_{n,2}}\right),$$

| | *Yeast* | | *Scene* | |
|---|---|---|---|---|
| | Accuracy | Log Likelihood | Accuracy | Log Likelihood |
| Jaakkola and Jordan (1996) | 79.7% | -0.678 | 87.4% | -0.670 |
| Laplace inference | **80.1%** | **-0.449** | **89.4%** | **-0.259** |

# More General $Q_\theta$

$$Q^* = Q_{\theta^*} : \theta^* = \arg\min_\theta D(Q_\theta(X), P(X|Z))$$

$X$ the latent variables and $Z$ the observations

- Gaussian Processes: A measure over continuous functions where any discrete sample of the domain follows a Gaussian law.

  $P(f(x)) : (f(x_1), \ldots, f(x_N)) \sim N(\mu_{x_1,\ldots,x_N}, \Sigma_{x_1,\ldots,x_N})$

- Support Transformations: $Q_\theta(X) \triangleq \phi_\theta(X)$

  $X \sim \mathcal{N}(\mu, \Sigma), \phi_\theta$ a parametric mass-preserving diffeomorphism

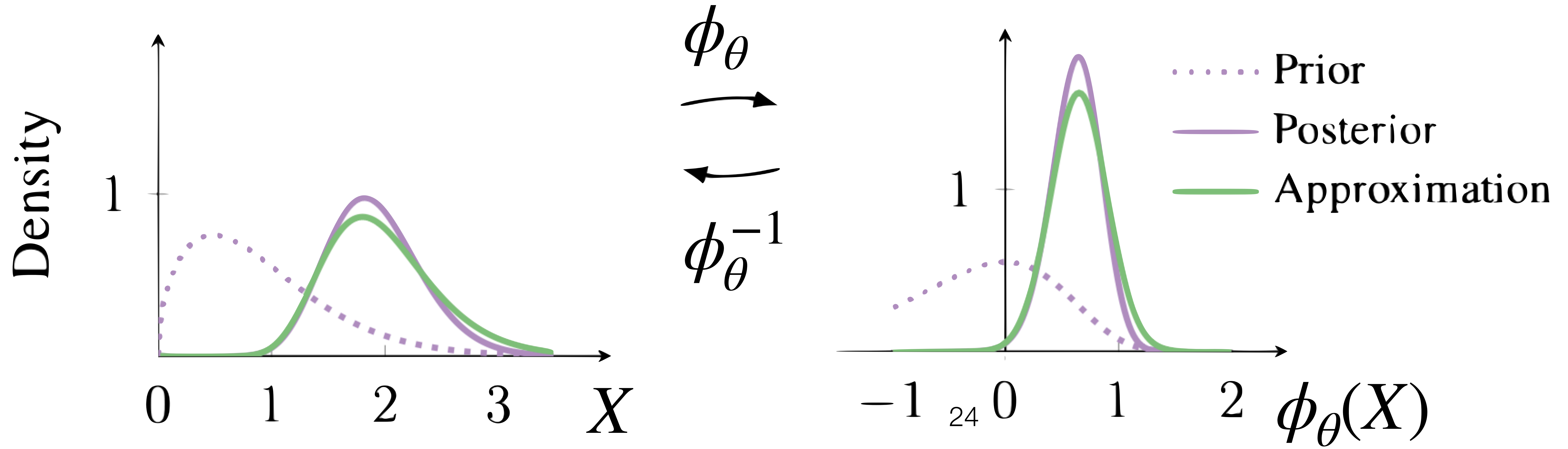
# More General $Q_\theta$

$$Q^* = Q_{\theta^*} : \theta^* = \arg\min_\theta D(Q_\theta(X), P(X\,|\,Z))$$

$X$ the latent variables and $Z$ the observations

- Support Transformations: $Q_\theta(X) \triangleq N_{\mu,\Sigma}(\phi_\theta(X))\left|J_{\phi_\theta}(X)\right|$

$X \sim \mathcal{N}(\mu, \Sigma)$, $\phi_\theta$ a parametric mass-preserving diffeomorphism




[Kucukelbir etal 17]

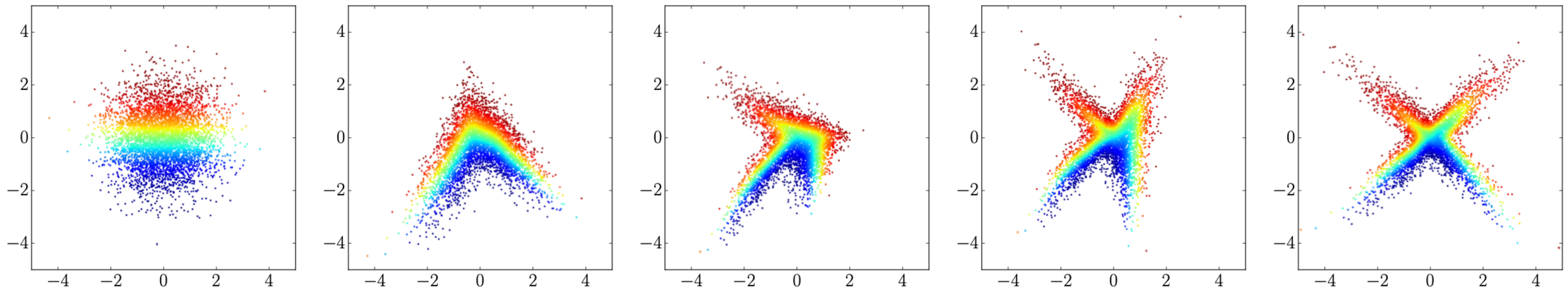
# More General $Q_\theta$

$$Q^* = Q_{\theta^*} : \theta^* = \arg\min_\theta D(Q_\theta(X), P(X \mid Z))$$

$X$ the latent variables and $Z$ the observations

- Support Transformations: $Q_\theta(X) \triangleq N_{\mu,\Sigma}(\phi_\theta(X)) \left| J_{\phi_\theta}(X) \right|$

$\phi_\theta(X) \sim \mathcal{N}(\mu, \Sigma)$, $\phi_\theta$ a stochastic flow or learnable diffeomorphism

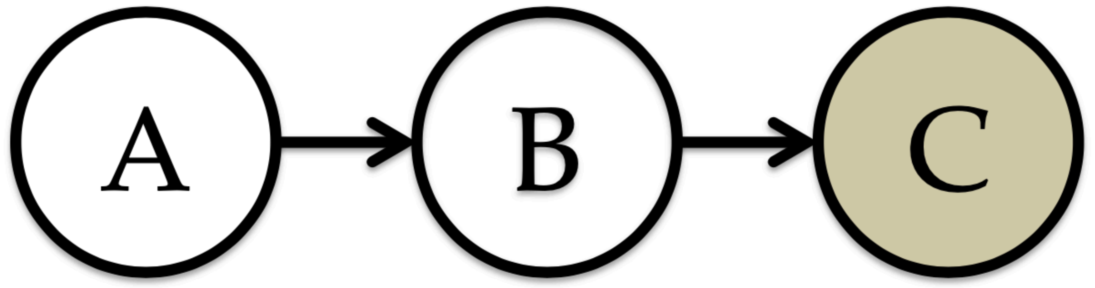





[Papamakarios etal 21]

# Current Problems in VI

- Scalability
- Amortization [Gershman et al 2014]
- Preservation of dependencies
- Auto-regressive models




Query 1: $P(B|C) = P(C|B)P(B)/P(C)$

Query 2: $P(A|C) = \sum_{B} P(A|B)P(B|C)$

Amortisation, reused probability in blue

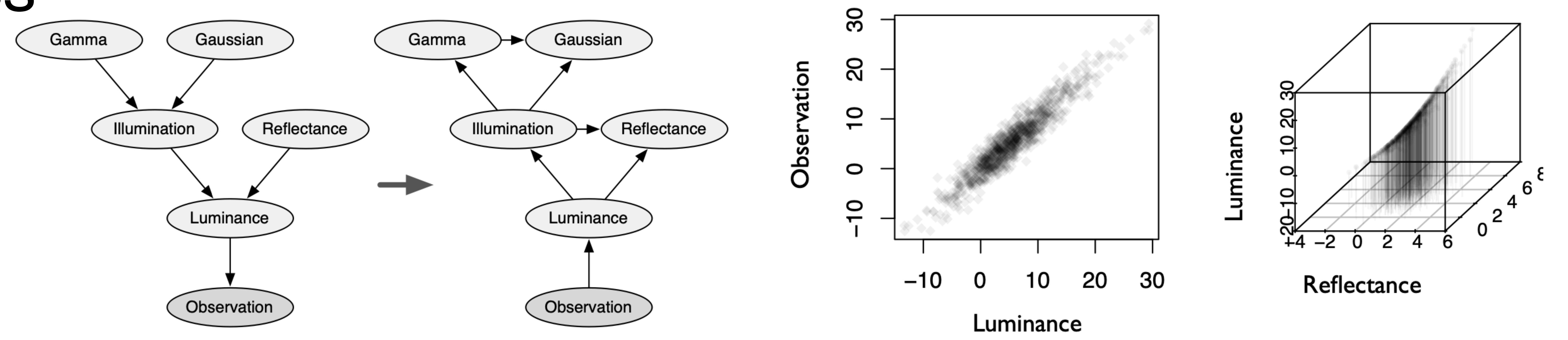


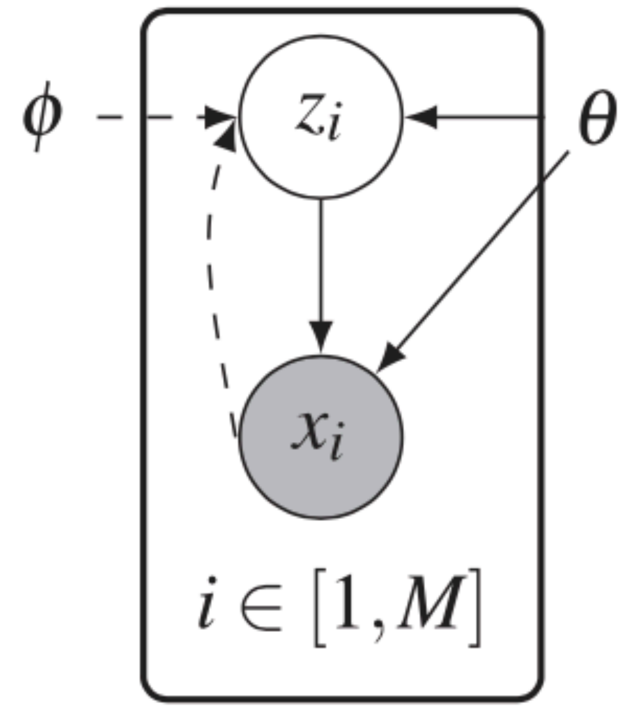

Figure 1: A Bayesian network modeling brightness constancy in visual perception, a possible invers
factorization, and two of the local joint distributions that determine the inverse conditionals.



[Stuhlmüller et al 14]

# Other Modern Bayesian Techniques

- Variational AutoEncoders
- Likelihood-free Inference



(b) VAE

$$\theta = \arg\max_{\theta} \mathscr{L}(\theta) = \mathbb{E}_{X \sim Q_\theta}\left[\log \frac{P(X,Z)}{Q_\theta(X)}\right]$$

$$VAE : Z \sim N(\mu(X), \Sigma(X))$$

Likelihood

Prior

$$P(Z|X) = \frac{P(X|Z)P(Z)}{P(X)}$$

Evidence