CSE 190 – Lecture 7

Data Mining and Predictive Analytics

Graphical Models

Gradesource

- You should have received an e-mail with your gradesource secret number
- Grades will be posted there once your homework has been marked
- Homework will be returned on Thursday (but I'll cover it in class next week)

Today

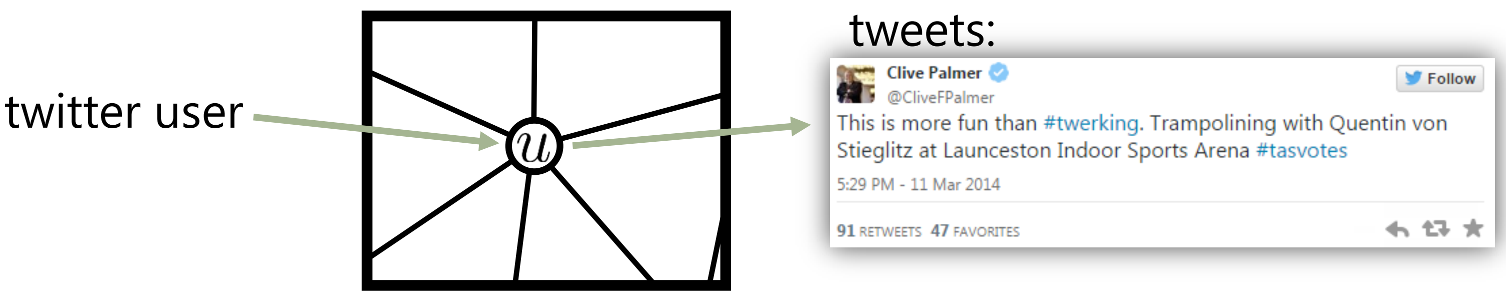
So far we've looked at prediction problems of the form

$$p(\text{label}|\text{data})$$

Today

e.g.

Estimate a user's political affiliation
from the content of their tweets



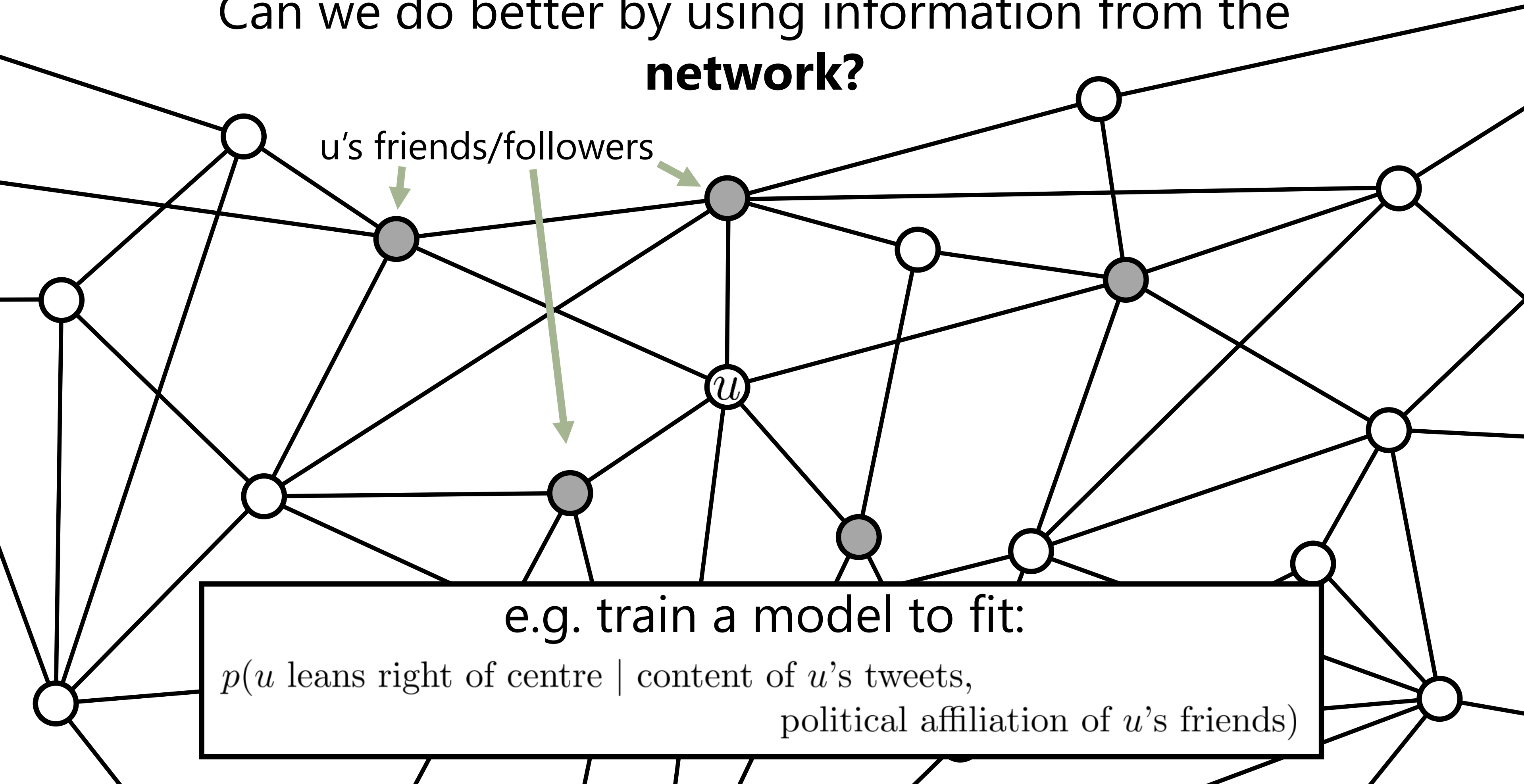
train a model to fit:

$$p(u \text{ leans right of centre} \mid \text{content of } u\text{'s tweets})$$

Today

But!

Can we do better by using information from the **network?**



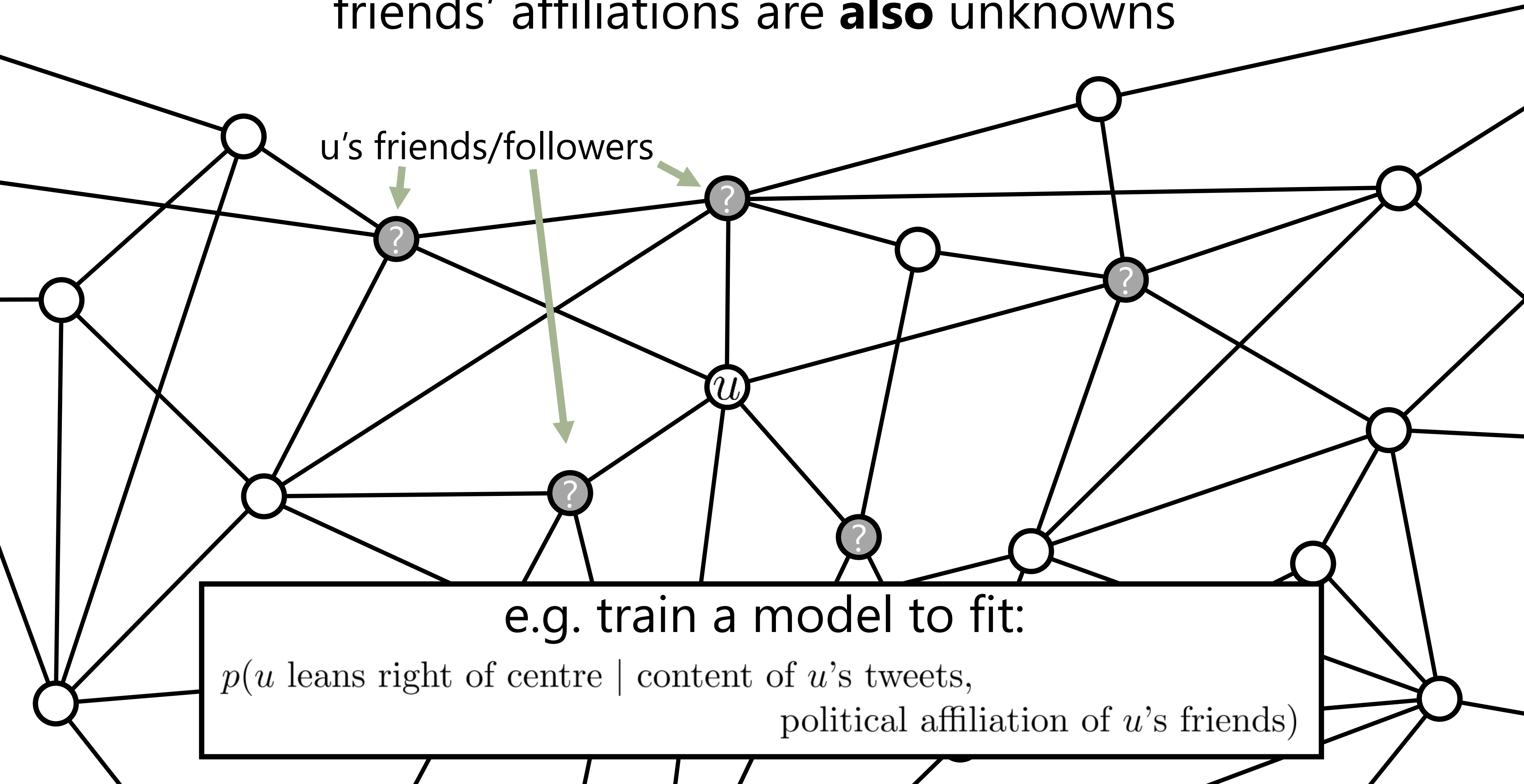
e.g. train a model to fit:

$$p(u \text{ leans right of centre} \mid \text{content of } u\text{'s tweets}, \\ \text{political affiliation of } u\text{'s friends})$$

Today

But (part 2)!

friends' affiliations are **also** unknowns



e.g. train a model to fit:

$$p(u \text{ leans right of centre} \mid \text{content of } u\text{'s tweets}, \\ \text{political affiliation of } u\text{'s friends})$$

Today

Interdependent variables

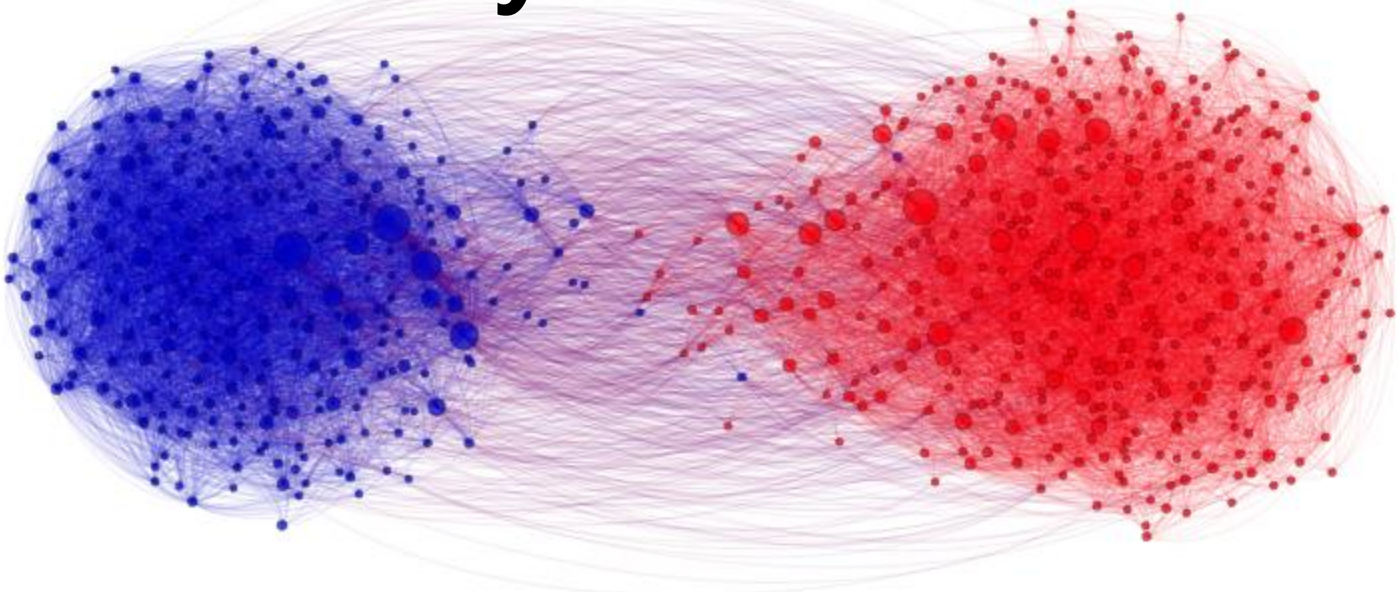
How can we solve predictive tasks when

- There are multiple unknowns to infer simultaneously
- There are **dependencies** between the unknowns
- In other words, what can we do when…

$$p(\text{label}_1, \text{label}_2 | \text{data}) \neq p(\text{label}_1 | \text{data}) p(\text{label}_2 | \text{data})$$

Examples

Infer the political affiliation of **every** user on twitter

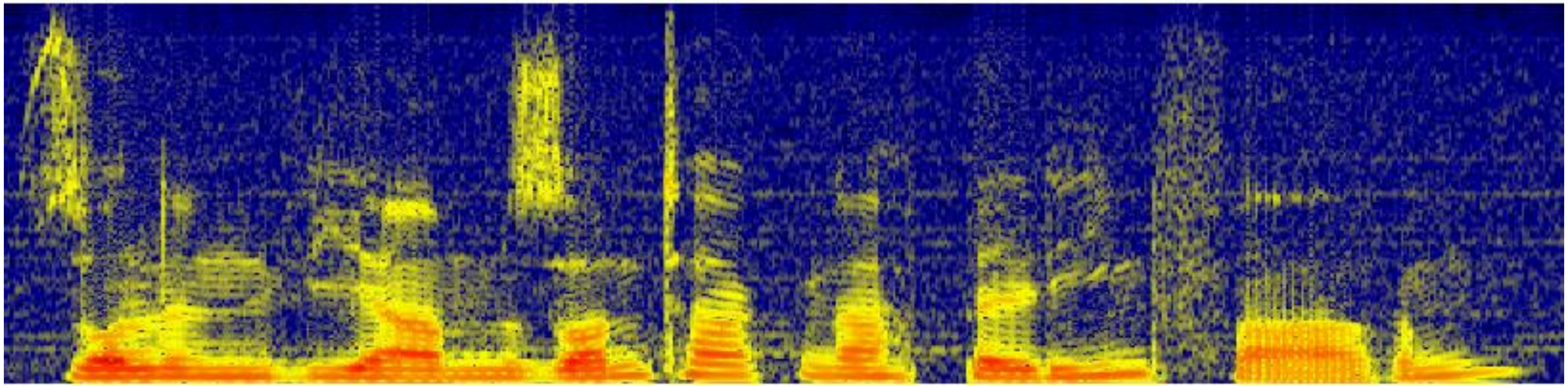


(kind of did this last week, but we didn't make any use of **evidence** at each node)

Graph data from Adamic (2004). Visualization from allthingsgraphed.com

Examples

What was said in the missing part of the signal?
(or, what was the **whole** signal)



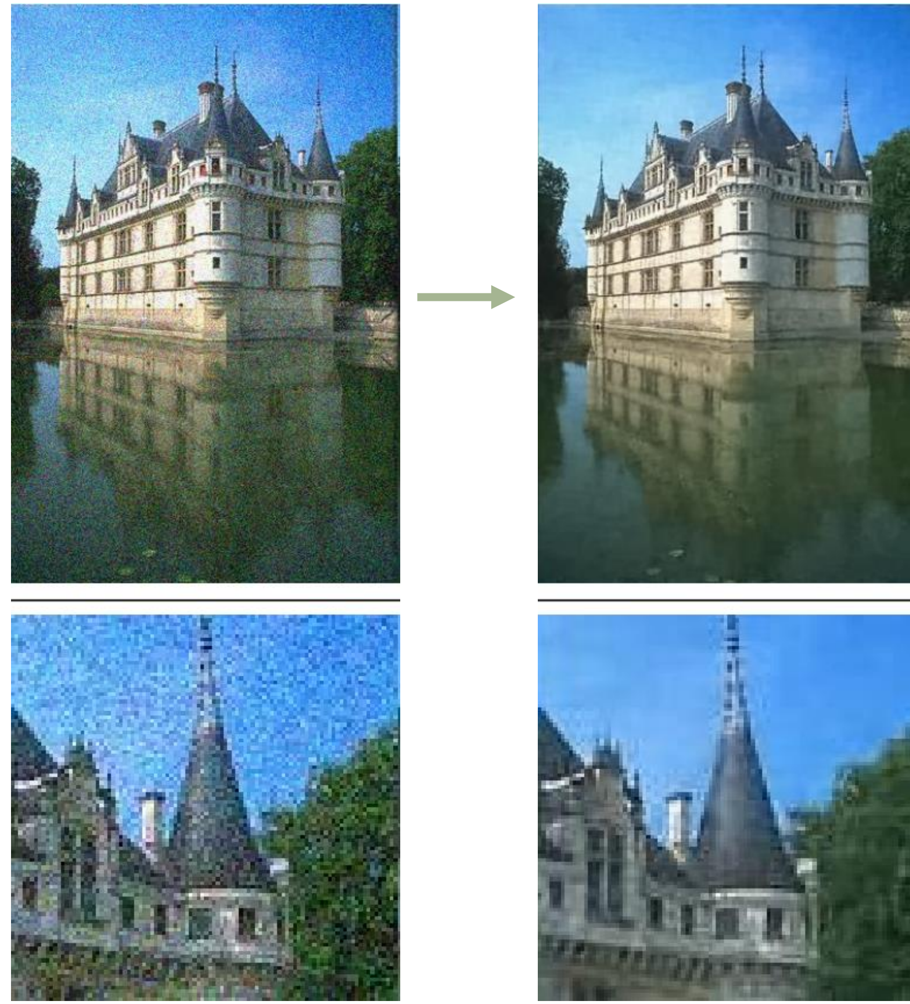
Sollen wir ? ?(garbled)? ? Berlin fahren

Image from http://www-i6.informatik.rwth-aachen.de/web/Research/speech_recog.html

Examples

Restore the image

The restored value of each pixel is related to (the restored value of) the pixels surrounding it



input output

Examples

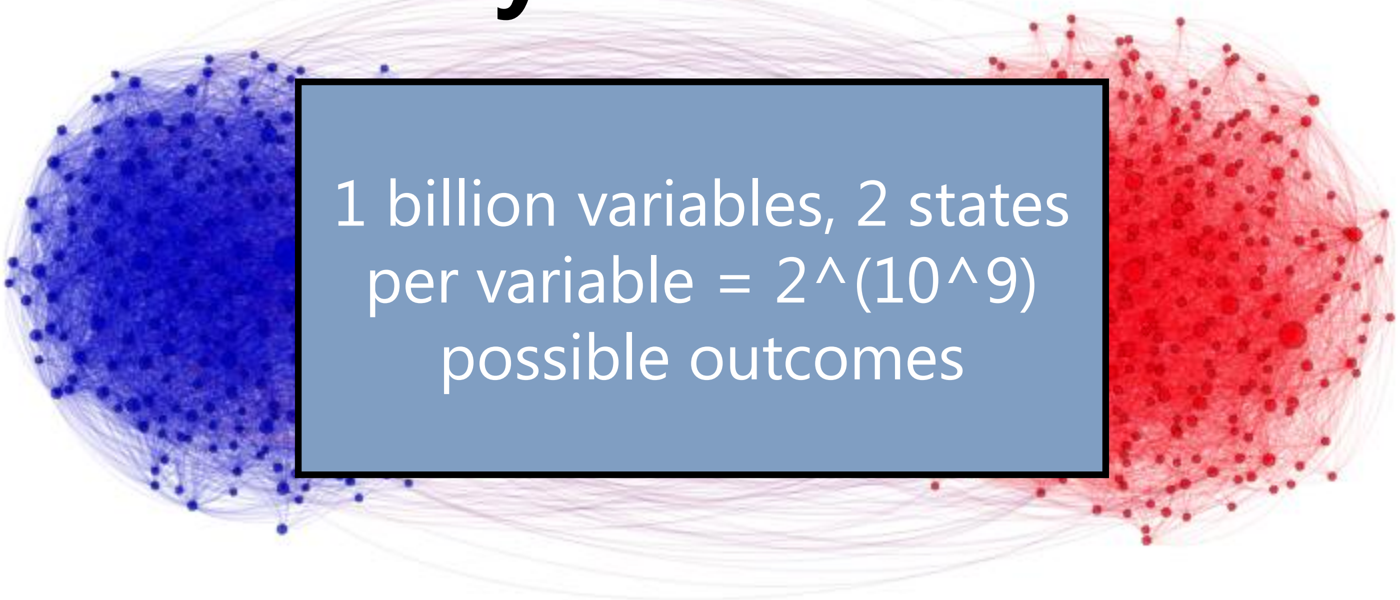
In all of these examples we can't infer the values of the unknown variables in isolation

(or at least not very well)

Q: Can we infer all of the variables **simultaneously** and account for their **interdependencies**?

Examples

Infer the political affiliation of **every** user on twitter

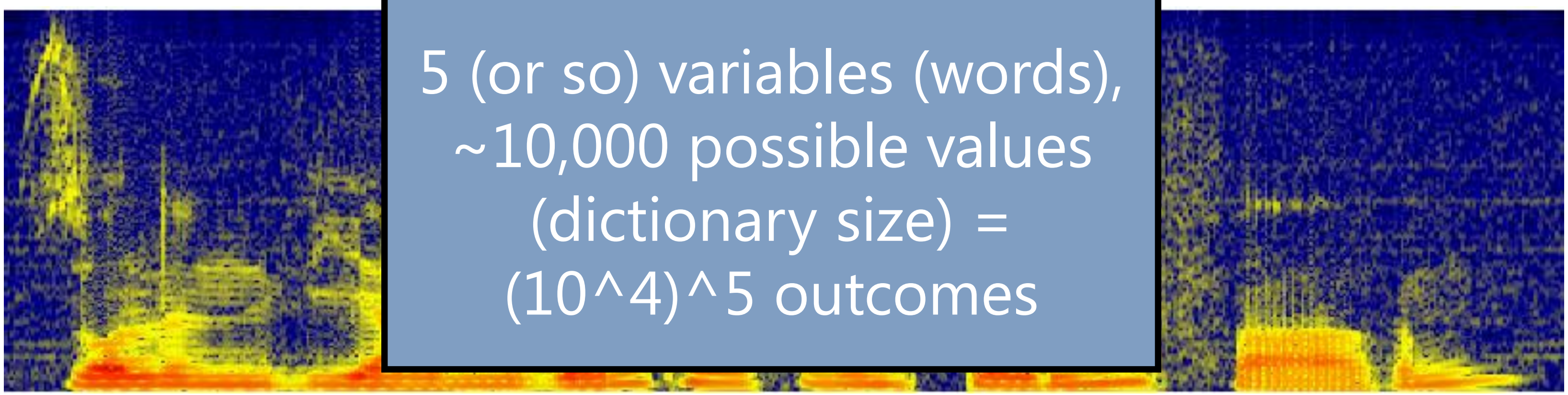


1 billion variables, 2 states per variable = 2^(10^9) possible outcomes

Graph data from Adamic (2004). Visualization from allthingsgraphed.com

Examples

What was said in the missing part of the signal?
(or, what was the **whole** signal)



5 (or so) variables (words),
~10,000 possible values
(dictionary size) =
(10^4)^5 outcomes

Sollen wir ? ?(garbled)? ? Berlin fahren

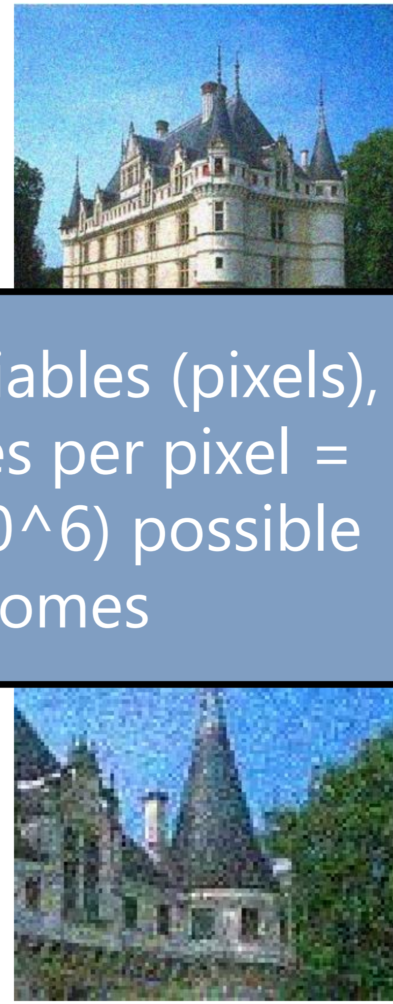
Image from http://www-i6.informatik.rwth-aachen.de/web/Research/speech_recog.html

Examples

Restore the image

The restored value of each pixel is related to (the restored value of surroundi

1 million variables (pixels), 256^3 states per pixel = (256^3)^(10^6) possible outcomes



input

output

Examples

A: State spaces are **way too big** to enumerate

But the problems are incredible **structured**, meaning that full enumeration may be avoidable

Examples

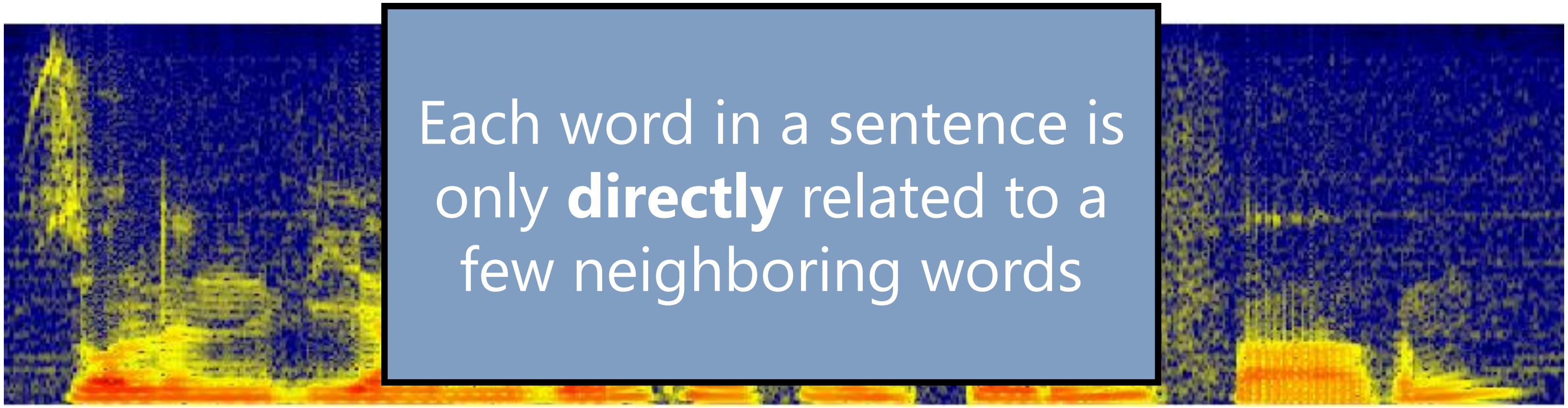
Infer the political affiliation of **every** user on twitter

My affiliation is only **directly** related to that of my friends

Graph data from Adamic (2004). Visualization from allthingsgraphed.com

Examples

What was said in the missing part of the signal?
(or, what was the **whole** signal)



Each word in a sentence is only **directly** related to a few neighboring words

Sollen wir ? ?(garbled)? ? Berlin fahren

Image from http://www-i6.informatik.rwth-aachen.de/web/Research/speech_recog.html

Examples

Restore the image

The restored value of each pixel is related to (the restored value of surroundi

Each pixel is only **directly** related to the few pixels nearby

input

output

Graphical models

Graphical models

- Are a **language** to describe the interdependencies between variables in multi-variable inference problems
- Give rise to a set of **algorithms** that exploit the structure of these interdependencies to make inference tractable

Today

- Some definitions
- Inference in chain-structured models
  (e.g. inference for sequence data)
- Inference in trees and networks that are "tree-like"
- Inference in some other useful and non-useful specific cases (maybe)
- Case study (maybe)

Probability distributions

Consider a high dimensional probability distribution such as:

$$p(a, b, c, d, e, f, g)$$

Such an expression can be rewritten as:

$$p(a)p(b|a)p(c|a,b)p(d|a,b,c)p(e|a,b,c,d)p(f|a,b,c,d,e)p(g|a,b,c,d,e,f)$$

Which is not so useful as it's still a function of seven variables, for example:

$$p(a) = \sum_{b,c,d,e,f,g} p(a, b, c, d, e, f, g)$$

is expensive to compute

Probability distributions

But what if a more useful factorization is possible?

$$p(a, b, c, d, e, f, g)$$

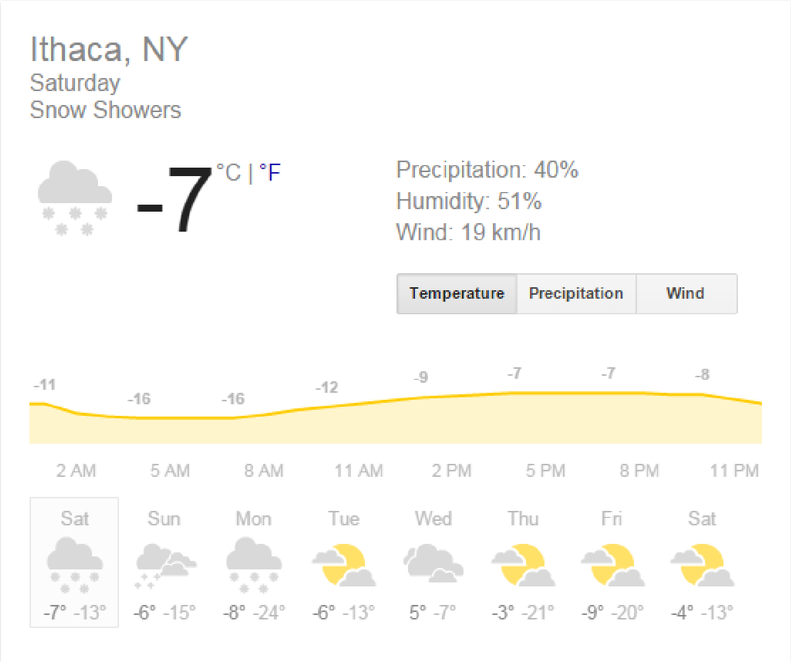
Imagine this can be rewritten as:

$$p(a)p(b|a)p(c|b)p(d|c)p(e|d)p(f|e)p(g|f)$$

"a causes b, b causes c, c causes d, d causes e…"

Probability distributions

e.g. what is the probability that the following forecast is accurate?



$$=p(Sun=-6 \mid Sat=-7)p(Mon=-8 \mid Sun=-6)p(Tue=-6 \mid Mon=-8)\ldots$$

Probability distributions

What is useful about a distribution that factorizes like

$$p(a)p(b|a)p(c|b)p(d|c)p(e|d)p(f|e)p(g|f)$$

is that we can compute marginals efficiently:

$$p(g) = \sum_{a,b,c,d,e,f} p(a,b,c,d,e,f,g)$$

$$= \sum_{a,b,c,d,e,f} p(a)p(b|a)p(c|b)p(d|c)p(e|d)p(f|e)p(g|f)$$

$$= \sum_{f} p(g|f) \sum_{e} p(f|e) \sum_{d} p(e|d) \sum_{c} p(d|c) \sum_{b} p(c|b) \sum_{a} p(a)p(b|a)$$

Probability distributions

$$p(a)p(b|a)p(c|b)p(d|c)p(e|d)p(f|e)p(g|f)$$

$$= \sum_f p(g|f) \underbrace{\sum_e p(f|e) \underbrace{\sum_d p(e|d) \underbrace{\sum_c p(d|c) \underbrace{\sum_b p(c|b) \underbrace{\sum_a p(a)p(b|a)}_{p(b) \quad O(N^2)}}_{p(c) \quad O(N^2)}}_{p(d) \quad O(N^2)}}_{p(e) \quad O(N^2)}}_{p(f) \quad O(N^2)}$$

(N = number of possible states per variable)

Probability distributions

We had a problem that was **expensive**:

$$p(g) = \sum_{a,b,c,d,e,f} p(a,b,c,d,e,f,g)$$

($O(N^K)$, N = number of states, K = number of variables)

but were able to solve it efficiently ($O(KN^2)$) due to factorization:

$$= \sum_f p(g|f) \sum_e p(f|e) \sum_d p(e|d) \sum_c p(d|c) \sum_b p(c|b) \underbrace{\sum_a p(a)p(b|a)}_{p(b)}$$

(bonus: we computed the marginal of every variable while we were at it!)

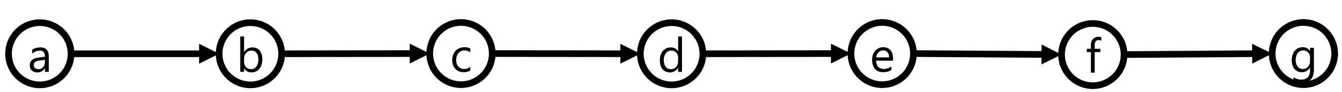
Directed graphical models (Bayes Nets)

Graphical models give us a language to describe such factorization assumptions e.g.

$$p(a, b, c, d, e, f, g)$$

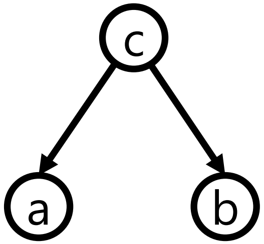
$$= p(a)p(b|a)p(c|b)p(d|c)p(e|d)p(f|e)p(g|f)$$

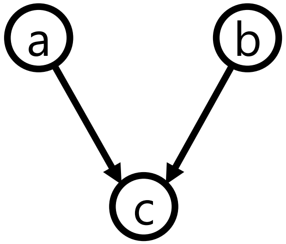
Can be described by the graph

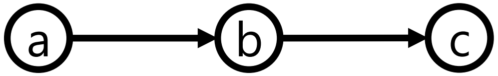


Directed graphical models

A few examples....







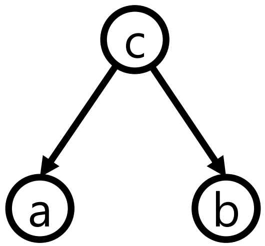
$p(c)p(a|c)p(b|c)$ $\quad$ $p(a)p(b)p(c|a,b)$ $\quad$ $p(a)p(b|a)p(c|b)$

Rule: terms factorize according to p(node|parents)

Directed graphical models

A few examples....

What about:



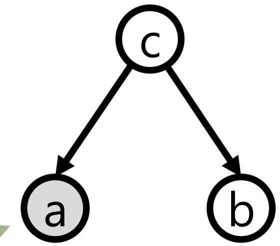
$$p(c)p(a|c)p(b|c)$$

What is: $p(b)$?

$$= \sum_{c,a} p(c)p(a|c)p(b|c)$$

$$= \sum_c p(c)p(b|c) \underbrace{\sum_a p(a|c)}_{= 1}$$

But what if we knew a?



evidence variable

$$p(c)p(a=1|c)p(b|c)$$

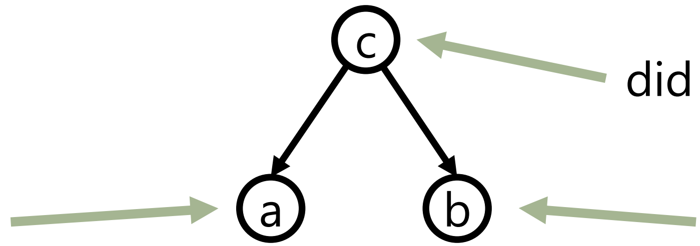
$p(b)$?

$$= \sum_c \underbrace{p(c)p(a=1|c)p(b|c)}_{= f(b,c)}$$

Conditional independence

What are the conditional independence statements implied by this graph?

are my knees grazed?



did I wreck my bicycle?

did I drive today?

$$p(c)p(a|c)p(b|c)$$

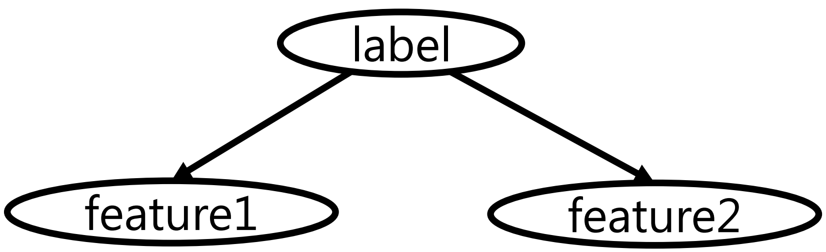
"*c* is a **common cause** for *a* and *b*"

"if we know *c*, then knowing *a* tells us nothing about *b*"

$$(a \perp\!\!\!\perp b|c)$$

Conditional independence

Recall: Naïve Bayes (week 2)

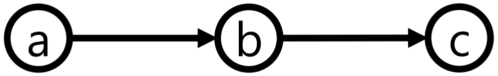


"features are independent given the label"

$$(feature_i \perp\!\!\!\perp feature_j | label)$$

Conditional independence

What are the conditional independence statements implied by this graph?



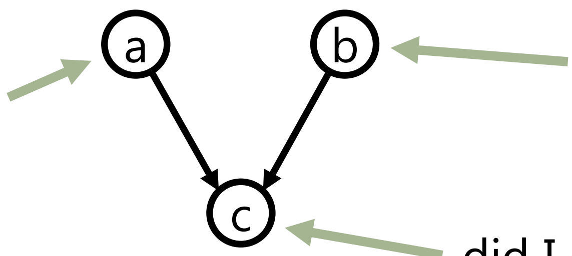
$$p(a)p(b|a)p(c|b)$$

$$(a \perp\!\!\!\perp c | b)$$

e.g. "Monday's weather is conditionally independent of Wednesday's weather, given Tuesday's weather"

Conditional independence

What are the conditional independence statements implied by this graph?



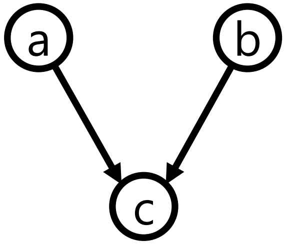
$$p(a)p(b)p(c|a,b)$$

$$(a \perp\!\!\!\perp b|c)?$$

No: e.g. think of a system with two points of failure.
If I know *c*, then knowing ~*a* tells me that *b* is likely.

Conditional independence

What are the conditional independence statements implied by this graph?



$$p(a)p(b)p(c|a,b)$$

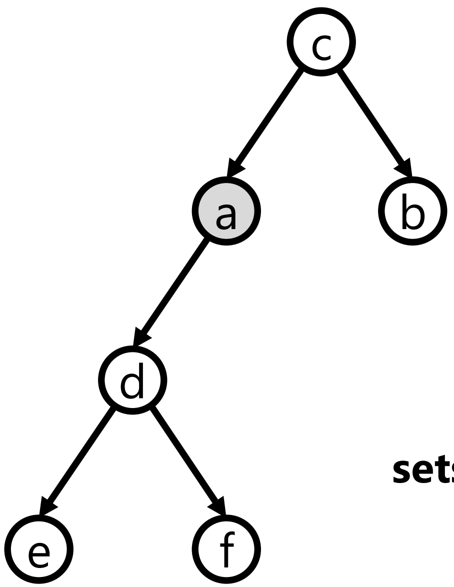
But… $p(a,b) = \sum_c p(a)p(b)p(c|a,b)$

$$(a \perp\!\!\!\perp b | \varnothing)$$

"*a* and *b* are conditionally independent if we know **nothing**"

D-separation

So… what parts of the graph can we ignore when doing inference?



e.g. if we know a, then we can ignore d,e,f when performing inference about b/c

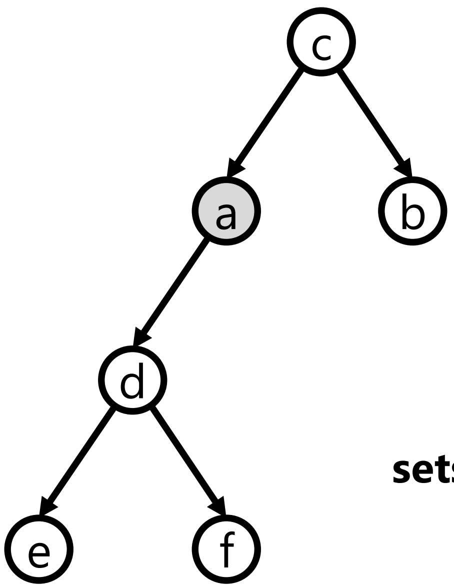
Case 1:

sets of nodes $(A \perp\!\!\!\perp B | C)$ if

any path from $a \in A$ to $b \in B$ meets at $\rightarrow c \rightarrow$ or $\leftarrow c \rightarrow$ with $c \in C$

D-separation

So… what parts of the graph can we ignore when doing inference?



e.g. if we know a, then we can ignore d,e,f when performing inference about b/c

Case 2:

sets of nodes $(A \perp\!\!\!\perp B | C)$ if

any path from $a \in A$ to $b \in B$ meets at $\rightarrow c \leftarrow$ and neither c **nor any of its decendants** are in C

D-separation

So… what parts of the graph can we ignore when doing inference?

In these two cases we say that C **d-separates** (directionally separates) A from B, and that $(A \perp\!\!\!\perp B | C)$

This means that if we know *C*, then we can ignore *B* when making inferences about *A*

These cases fully characterize the independence structure of the distribution (Pearl, 1988)

Questions

Further reading:

- Bishop Chapter 8
- Coursera course on PGMs: https://www.coursera.org/course/pgm
- More on d-separation (from the source) – Geiger, Verma, & Pearl, 1990: http://ftp.cs.ucla.edu/pub/stat_ser/r116.pdf

CSE 190 – Lecture 7

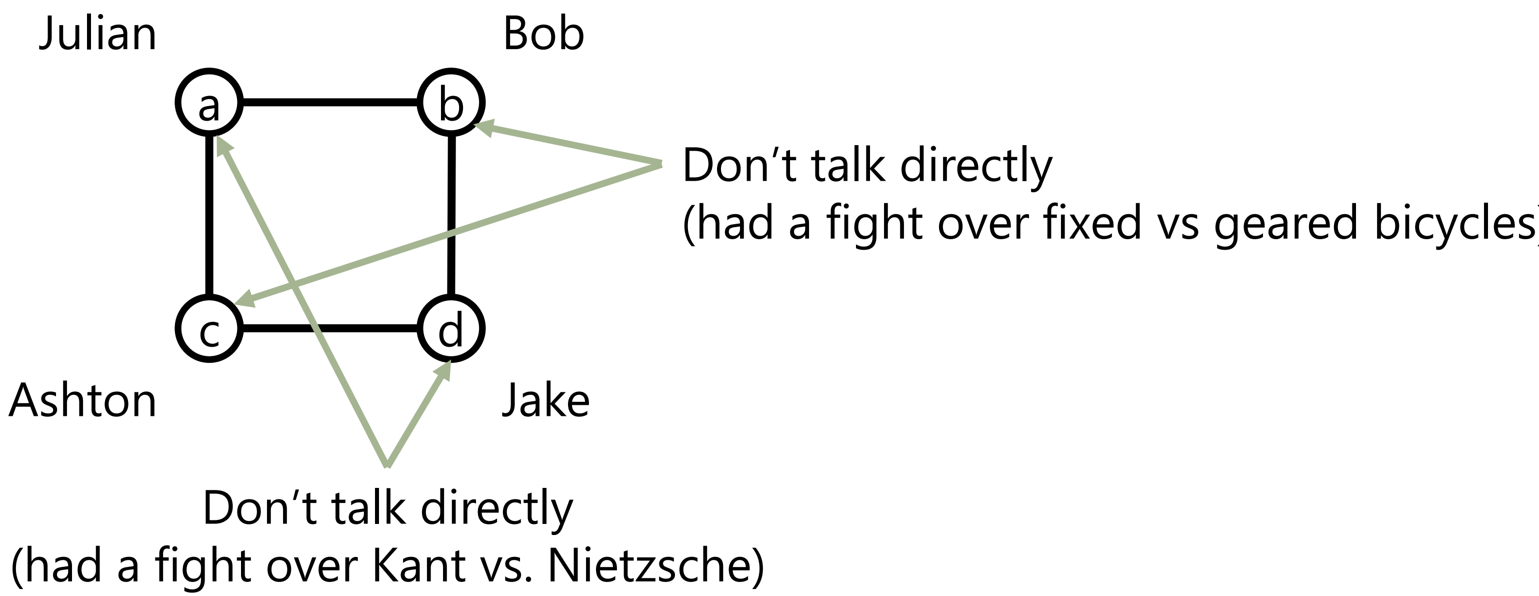
Data Mining and Predictive Analytics

Undirected Graphical Models

Undirected graphical models

Consider the following social network:

(in which friends influence each other's decisions)

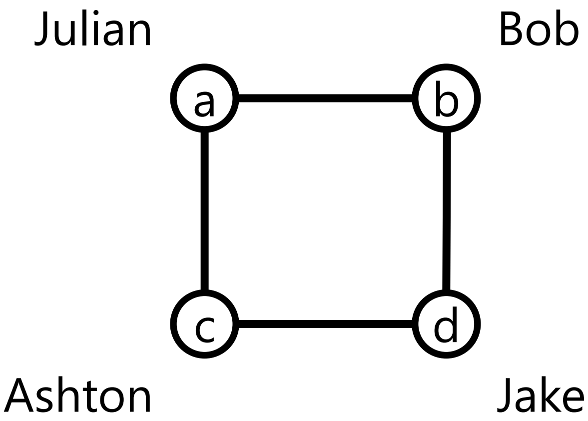


Who will vote the same way?

(see similar examples in slides from Stanford (Koller), Buffalo (Srihari) etc.)

Undirected graphical models

What graphical model represents this?



Want:

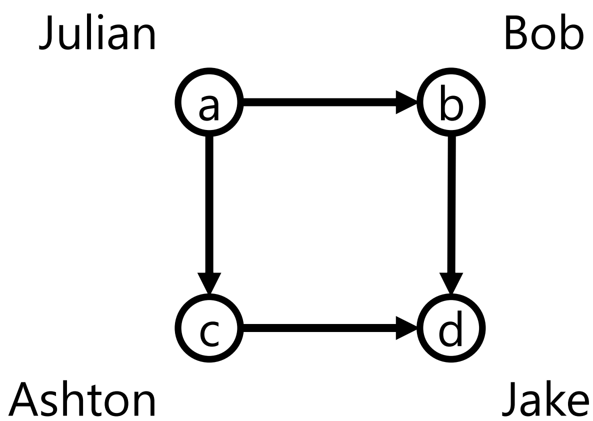
$$(julian \perp\!\!\!\perp jake | ashton, bob)$$
$$(a \perp\!\!\!\perp d | b, c)$$

$$(ashton \perp\!\!\!\perp bob | julian, jake)$$
$$(b \perp\!\!\!\perp c | a, d)$$

Who will vote the same way?

Undirected graphical models

Attempt 1:



Want:

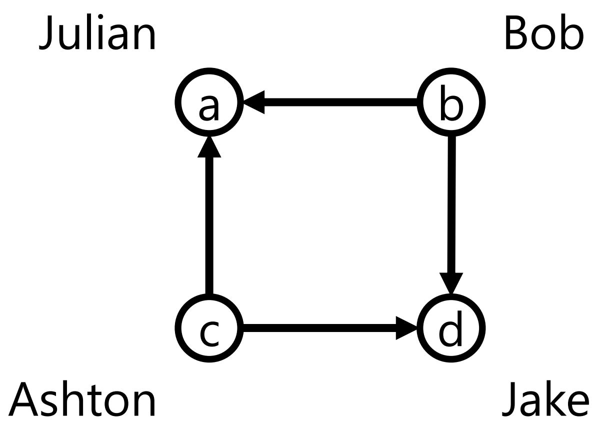
$$(julian \perp\!\!\!\perp jake | ashton, bob)$$
$$(a \perp\!\!\!\perp d | b, c)$$ **yes**

$$(ashton \perp\!\!\!\perp bob | julian, jake)$$
$$(b \perp\!\!\!\perp c | a, d)$$ **no (why?)**

Who will vote the same way?

Undirected graphical models

Attempt 2:



Want:

$$(julian \perp\!\!\!\perp jake | ashton, bob)$$
$$(a \perp\!\!\!\perp d | b, c)$$

yes

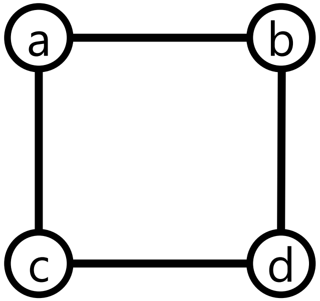
$$(ashton \perp\!\!\!\perp bob | julian, jake)$$
$$(b \perp\!\!\!\perp c | a, d)$$

no (why?)

Who will vote the same way?

Undirected graphical models

There **is no** directed network that will capture exactly these conditional independence assumptions

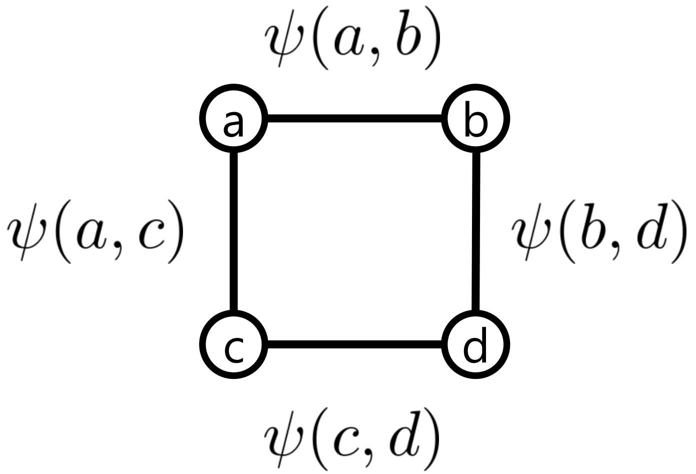


$$(a \perp\!\!\!\perp d | b, c)$$
$$(b \perp\!\!\!\perp c | a, d)$$

So let's use an undirected network to represent them!

Undirected graphical models

The edges of the network determine how the distribution **factorizes**

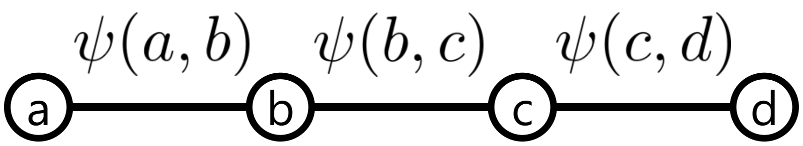


$$p(a, b, c, d) = \frac{1}{Z}\psi(a, b)\psi(a, c)\psi(b, d)\psi(c, d)$$

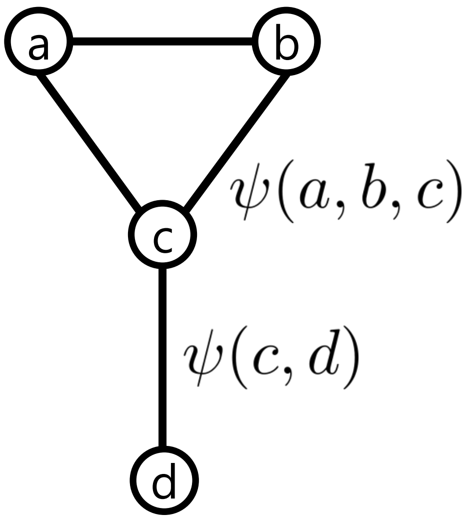
$$Z = \sum_{a,b,c,d} \psi(a, b)\psi(a, c)\psi(b, d)\psi(c, d)$$

Undirected graphical models

Examples:



$$p(a,b,c,d) = \frac{1}{Z}\psi(a,b)\psi(b,c)\psi(c,d)$$



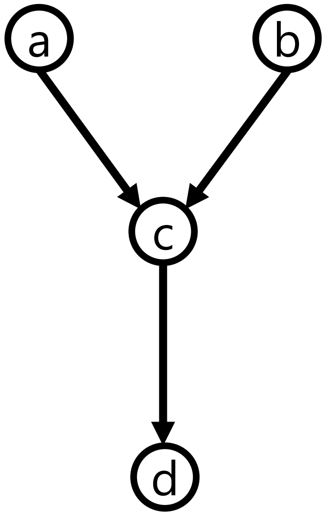
$$p(a,b,c,d) = \frac{1}{Z}\psi(a,b,c)\psi(c,d)$$

Factors are defined over the (maximal) **cliques** of the graph

Undirected graphical models

How to convert from a directed to an undirected network?

e.g.

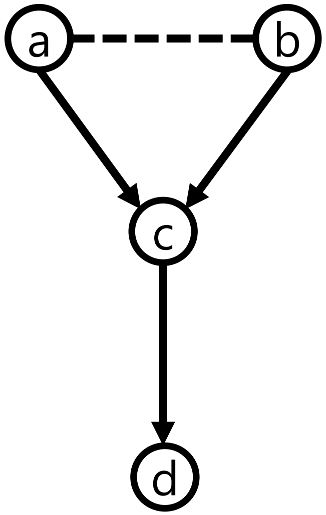


$$p(a, b, c, d) = p(a)p(b)p(c|a, b)p(d|c)$$

Undirected graphical models

How to convert from a directed to an undirected network?

e.g.



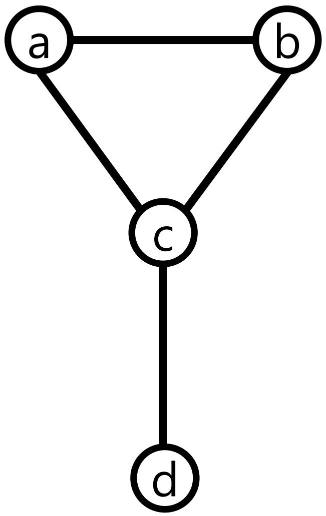
(1)
connect the
parents
of each node
("moralization")

$$p(a, b, c, d) = p(a)p(b)p(c|a, b)p(d|c)$$

Undirected graphical models

How to convert from a directed to an undirected network?

e.g.



(2)
disregard edge directedness

$$p(a,b,c,d) = \underbrace{p(a)p(b)p(c|a,b)}\underbrace{p(d|c)}$$

$$p(a,b,c,d) = \frac{1}{Z}\psi(a,b,c)\psi(c,d)$$

Undirected graphical models

How to convert from a directed to an undirected network?

every term from the original
graph now appears in a clique

$$p(a,b,c,d) = \underbrace{p(a)p(b)p(c|a,b)}\underbrace{p(d|c)}$$

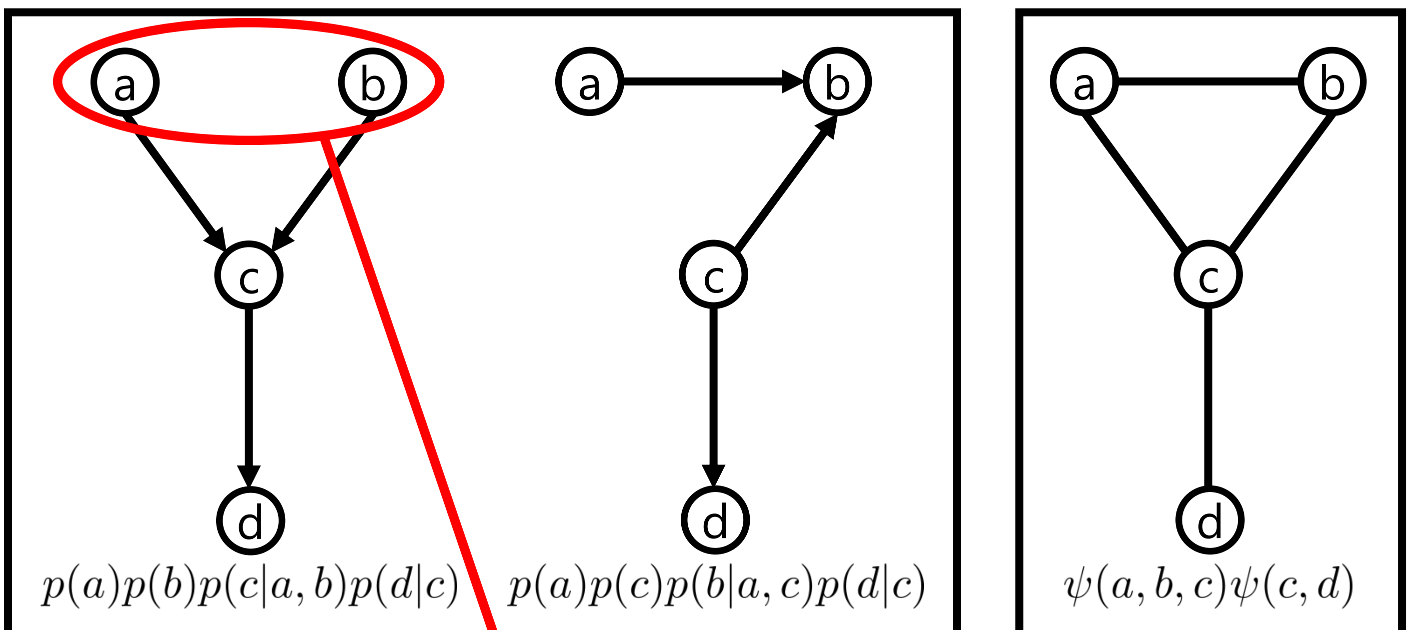
$$p(a,b,c,d) = \frac{1}{Z}\psi(a,b,c)\psi(c,d)$$

But: the construction has
"forgotten" some information

Undirected graphical models

How to convert from a directed to an undirected network?

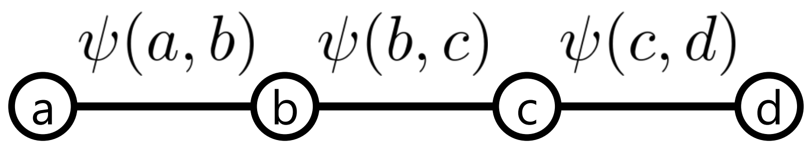
Both directed networks transform to the same undirected network



But we lost the fact that $(a \perp\!\!\!\perp b|\varnothing)$ in the undirected version

Undirected graphical models

Inference is similar to the directed case:



$$p(a,b,c,d) = \frac{1}{Z}\psi(a,b)\psi(b,c)\psi(c,d)$$

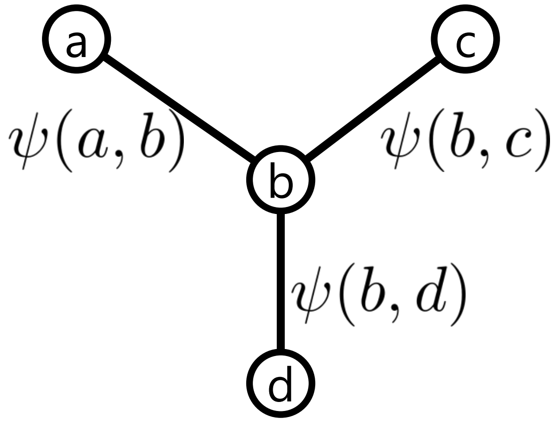
$$p(a) = \sum_{b,c,d} \frac{1}{Z}\psi(a,b)\psi(b,c)\psi(c,d)$$

$$= \frac{1}{Z}\sum_b \psi(a,b) \sum_c \psi(b,c) \sum_d \psi(c,d)$$

Just normalize the result so that it's a probability distribution

Undirected graphical models

Another example...



$$p(a) = \sum_{b,c,d} \frac{1}{Z} \psi(a,b)\psi(b,c)\psi(b,d)$$

$$= \frac{1}{Z} \sum_b \psi(a,b) \sum_c \psi(b,c) \sum_d \psi(b,d)$$

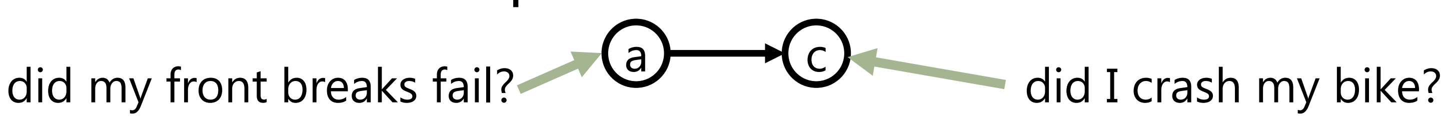
multiple elimination orderings are possible:

$$= \frac{1}{Z} \sum_b \psi(a,b) \sum_d \psi(b,d) \sum_c \psi(b,c)$$

Where are we so far?

We have a “language” for describing the dependencies between variables in multi-variable inference problems

- **Directed** graphical models are a natural way to represent causal relationships, e.g. whether I drive to work depends on whether it rains



- **Undirected** graphical models are a natural way to represent mutual dependencies, e.g. friends are likely to share similar opinions

What do we still need to do?

We still need to use these models to perform inference tasks (i.e., to infer marginals and maximum likelihood states)

- So far we've done this by moving around summation signs on the blackboard until things worked
- But how can we define **algorithms** to do this **automatically?**
- See lecture supplement on course webpage

Questions?

Next lecture:

- Homework 1
- Some recap of midterm-related material, and any pressing questions
- (maybe) get started on recommender systems