

# Bayesian Networks – Representation

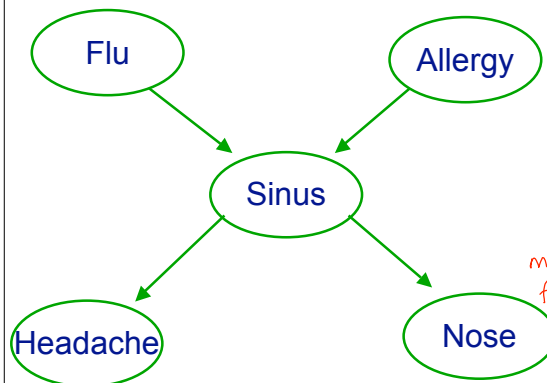
Machine Learning – 10701/15781  
Carlos Guestrin  
Carnegie Mellon University

October 31<sup>st</sup>, 2007

©2005-2007 Carlos Guestrin

1

## Possible queries



### ■ Inference

$$P(F=t \mid H=t, N=f)$$

### ■ Most probable explanation

$$\max_{f, a, s} P(f, a, s \mid H=t, N=f)$$

### ■ Active data collection

what should I measure

©2005-2007 Carlos Guestrin

2

# Factored joint distribution - Preview

Notation  $F, A \rightarrow$  I am not specify an assignment  
 $f, a \rightarrow$  specific assignments  
 $F=t \rightarrow \text{Flu} = \text{true}$  (a particular assignment)

$P(F)$   
 $P(A)$   
 $P(S|F,A)$   
 $P(H|S)$   
 $P(N|S)$

$P(F, A, S, H, N)$   
 $\uparrow 2^5 - 1$  (because sums to 1)  
 $r = 32 - 1 = 31$

$P(F, A, S, H, N) = P(F) \cdot P(A) \cdot P(S|F,A) \cdot P(H|S) \cdot P(N|S)$

$P(F) =$

$t$	0.1
$f$	0.9

$P(H|S) :$

$S$	$t$	$f$
$t$	0.8	0.3
$f$	1-0.8 0.2	0.7

2 numbers

©2005-2007 Carlos Guestrin 3

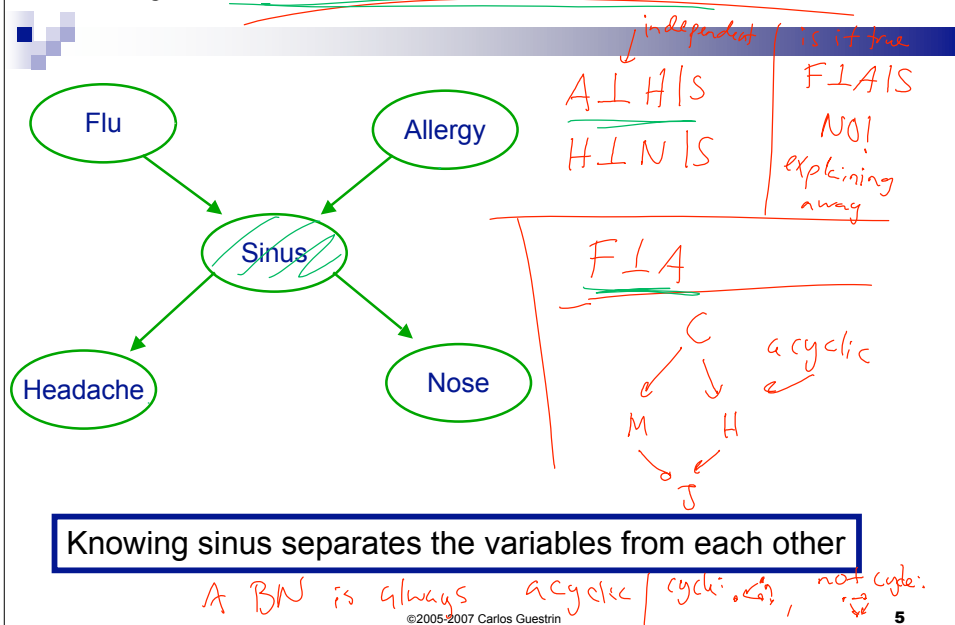
# Number of parameters

$P(F) \in 1 \text{ param}$   
 $P(A) \in 1$   
 $P(S|F,A) \in 4 \text{ params}$   
 $P(H|S) \in 2$   
 $P(N|S) \in 2$

total: 10

©2005-2007 Carlos Guestrin 4

# Key: Independence assumptions



## (Marginal) Independence

- Flu and Allergy are (marginally) independent

$F \perp A$

$P(F, A) = P(F) \cdot P(A)$

- More Generally:

$P(F|A) = P(F)$

Flu = t	0.2
Flu = f	0.8

Allergy = t	0.3
Allergy = f	0.7

	Flu = t	Flu = f
Allergy = t	$0.3 \times 0.2$	$0.3 \times 0.8$
Allergy = f	$0.2 \times 0.7$	$0.7 \times 0.8$

# Marginally independent random variables

- Sets of variables  $X, Y$
- X is independent of Y if  $\forall x \in \text{Val}(X), y \in \text{Val}(Y)$ 
  - ~~$P(X=x, Y=y) = P(X=x) \cdot P(Y=y)$~~ ,  ~~$x \in \text{Val}(X), y \in \text{Val}(Y)$~~   
 $P(X=x, Y=y) = P(X=x) \cdot P(Y=y)$
- Shorthand:  $P(X=x | Y=y) = P(X=x)$ 
  - Marginal independence:  ~~$X \perp Y$~~  ( $X \perp Y$ )
- Proposition:  $P$  satisfies  $(X \perp Y)$  if and only if
  - $P(X, Y) = P(X) P(Y)$   
 $P(X | Y) = P(X)$

# Conditional independence

- Flu and Headache are not (marginally) independent  
 $P(F | H) \neq P(F)$
- Flu and Headache are independent given Sinus infection  
 $P(H | S) = P(H | S, F)$   $F \perp H | S$   
 $P(H=t) = 0.1$   $\Downarrow$   $P(H=t | S=t, F=t) = 0.7$   
 $P(H=t | S=t) = 0.7$
- More Generally:  $X_i \perp X_j | X_k$   
 $P(X_i | X_j, X_k) = P(X_i | X_k)$   
 $P(X_i, X_j | X_k) = P(X_i | X_k) \cdot P(X_j | X_k)$

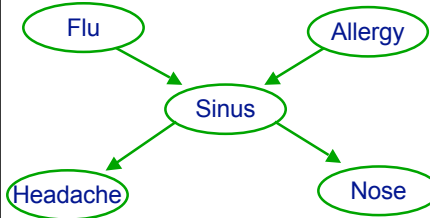
## Conditionally independent random variables

- Sets of variables  $X, Y, Z$
- X is independent of Y given Z if
  - $P(X=x \perp Y=y | Z=z), \forall x \in \text{Val}(X), \forall y \in \text{Val}(Y), \forall z \in \text{Val}(Z)$   
 $P(X=x | Y=y, Z=z) = P(X=x | Z=z)$
- Shorthand:
  - **Conditional independence:**  $X \perp Y | Z$
  - For  $(X \perp Y | Z)$ , write  $X \perp Y$
- **Proposition:**  $P$  satisfies  $(X \perp Y | Z)$  if and only if
  - $P(X, Y | Z) = P(X | Z) P(Y | Z)$

## Properties of independence

- **Symmetry:**
  - $(X \perp Y | Z) \Rightarrow (Y \perp X | Z)$
- **Decomposition:**
  - $(X \perp Y, W | Z) \Rightarrow (X \perp Y | Z)$
- **Weak union:**
  - $(X \perp Y, W | Z) \Rightarrow (X \perp Y | Z, W)$
- **Contraction:**
  - $(X \perp W | Y, Z) \& (X \perp Y | Z) \Rightarrow (X \perp Y, W | Z)$
- **Intersection:**
  - $(X \perp Y | W, Z) \& (X \perp W | Y, Z) \Rightarrow (X \perp Y, W | Z)$
  - Only for positive distributions!
  - $P(\alpha) > 0, \forall \alpha, \alpha \neq;$

# The independence assumption



**Local Markov Assumption:**  
 A variable  $X$  is independent of its non-descendants given its parents

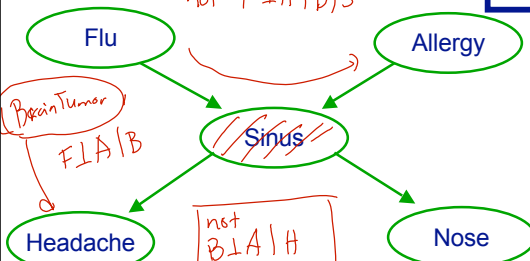
$F \perp A \mid \emptyset$   
 $S \perp \{F, A\}$  ? *noting*  


---

 $H \perp \{F, A, N\} \mid S$   
 $N \perp \{H, F, A\} \mid S$

$F \perp A$   
 $F \perp H \mid S$   
 $A \perp N \mid S$   
 $H \perp N \mid S$  ..

# Explaining away



**Local Markov Assumption:**  
 A variable  $X$  is independent of its non-descendants given its parents *and only its parents*

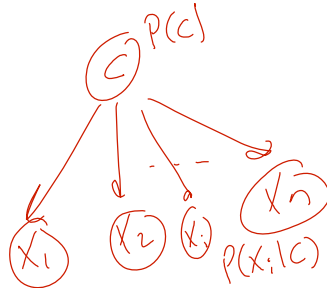
*not  $F \perp A \mid B, S$*   
 $F \perp A$  (jumping ahead)  
 $\downarrow$  *not:  $F \perp A \mid N$*   
 $F \perp A \mid S \leftarrow \text{NO!!}$   
 $P(F=t) = 0.2$ ,  $P(F=t \mid A=t) = 0.2 = P(F=t)$   
 $P(F=t \mid S=t) = 0.7 \neq P(F=t \mid S=t, A=t) = 0.4$   
*not:  $F \perp A \mid S$*  why? Allergies help explain  $S=t$   
*goes down!*

# Naïve Bayes revisited

$$P(C, x_1, \dots, x_n) = P(C) \cdot \prod_i P(x_i | C)$$

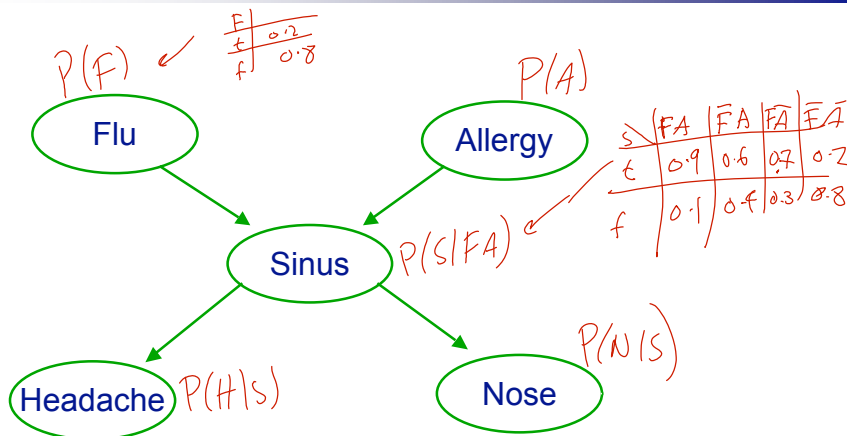
*X<sub>i</sub>'s are independent given C*

**Local Markov Assumption:**  
A variable X is independent of its non-descendants given its parents

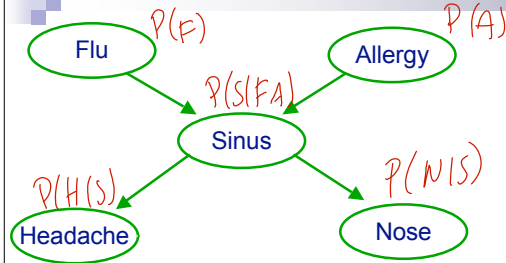


$$X_i \perp \{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n\} | C$$

# What about probabilities? Conditional probability tables (CPTs)



## Joint distribution



$$P(F, A, S, H, N) = P(F) \cdot P(A) \cdot P(S|FA) \cdot P(H|S) \cdot P(N|S)$$

**Why can we decompose? Markov Assumption!**

## The chain rule of probabilities

■  $P(A, B) = P(A)P(B|A)$  *exactly (no assumptions)*

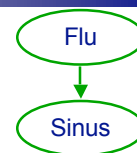
$$P(F, S) = P(F) \cdot P(S|F)$$

$$P(S|F) = \frac{P(F, S)}{P(F)}$$

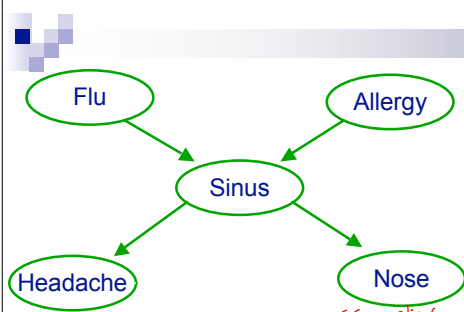
■ More generally:

$$\square P(X_1, \dots, X_n) = P(X_1) \cdot P(X_2|X_1) \cdot \dots \cdot P(X_n|X_1, \dots, X_{n-1})$$

$$P(X_1, \dots, X_n) = P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_2, X_1) \cdot \dots$$



# Chain rule & Joint distribution



**Local Markov Assumption:**  
A variable X is independent of its non-descendants given its parents

no assumptions

$$P(F, A, S, H, N) = P(F) \cdot P(A|F) \cdot P(S|FA) \cdot P(H|FAS) \cdot P(N|FASH)$$

with Markov Assumption

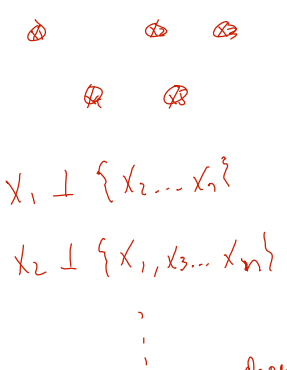
$$= P(F) P(A) \cdot P(S|FA) \cdot P(H|S) \cdot P(N|S)$$


---

$P(A|F) = P(A)$       $P(H|F,AS) = P(H|S)$       $P(N|F,AS,H) = P(N|S)$   
 $A \perp F$                       $H \perp \{F,AS\}$                       $N \perp \{A,H,F\}|S$

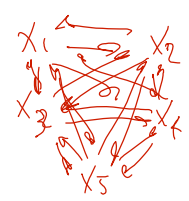
# Two (trivial) special cases

## Edgeless graph



All possible independencies  
lots of bias

## Fully-connected graph



no assumptions!!

no bias  
can represent any dist.

# The Representation Theorem – Joint Distribution to BN

*represent  $\equiv$  perfectly represent*

**BN:**  **Encodes independence assumptions**

If conditional independencies in BN are subset of conditional independencies in  $P$

**Obtain** 

**Joint probability distribution:**

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | \text{Pa}_{X_i})$$

*you can represent P with this BN*

*the more independencies in the BN, the fewer P's you can represent. But the easier it will be to learn the BN*

©2005-2007 Carlos Guestrin

19

## A general Bayes net

- Set of random variables  $X_1, \dots, X_n$   
*F, A, H, ...*
- Directed acyclic graph
  - Encodes independence assumptions



- CPTs  $P(X_i | \text{Pa}_{X_i})$ , e.g.,  $P(S|FA)$   
 $P(H|S)$   
...

- Joint distribution:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | \text{Pa}_{X_i})$$

*$P(A) \cdot P(F) \cdot P(S|FA) \dots$*

# How many parameters in a BN?

- Discrete variables  $X_1, \dots, X_n$  binary  
k-ary
  - Graph
    - Defines parents of  $X_i$ ,  $\text{Pa}_{X_i}$
  - CPTs –  $P(X_i | \text{Pa}_{X_i})$   $|\text{val}(X_i)| = k$
- $P(X_1, \dots, X_n) \in 2^n - 1$  params
- $k^n - 1$
- for each combination of parents  $\times$  # assignments of  $X_i - 1$
- $k^{|\text{Pa}_{X_i}|} \times (k-1)$
- (X)  $P(X_1)$  (X)  $P(X_2)$  k-1 params

(X)  $P(X_3)$
- # params in a BN  
 $= \sum_i (k-1) k^{|\text{Pa}_{X_i}|}$
- # param in ~~directed~~ graph  
 $n \times (k-1)$

# Real Bayesian networks applications

- Diagnosis of lymph node disease
- Speech recognition
- Microsoft office and Windows
  - <http://www.research.microsoft.com/research/dtg/>
- Study Human genome
- Robot mapping
- Robots to identify meteorites to study
- Modeling fMRI data
- Anomaly detection
- Fault dianosis
- Modeling sensor network data



# Independencies encoded in BN

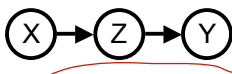
- We said: All you need is the local Markov assumption
  - $(X_i \perp \text{NonDescendants}_{X_i} \mid \text{Pa}_{X_i})$ , and ~~only~~ *only*  $\text{Pa}_{X_i}$
- But then we talked about other (in)dependencies
  - e.g., explaining away
  
- What are the independencies encoded by a BN?
  - Only assumption is local Markov
  - But many others can be derived using the algebra of conditional independencies!!!

## Understanding independencies in BNs

### – BNs with 3 nodes

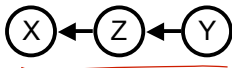
**Local Markov Assumption:**  
A variable X is independent of its non-descendants given its parents

Indirect causal effect:



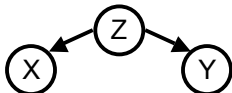
$Y \perp X \mid Z$

Indirect evidential effect:



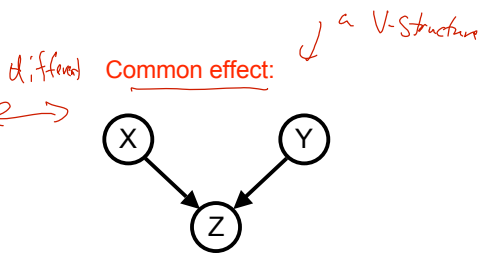
$Y \perp X \mid Z$

Common cause:



$Z \perp X \mid Y$

*equivalent!!!*

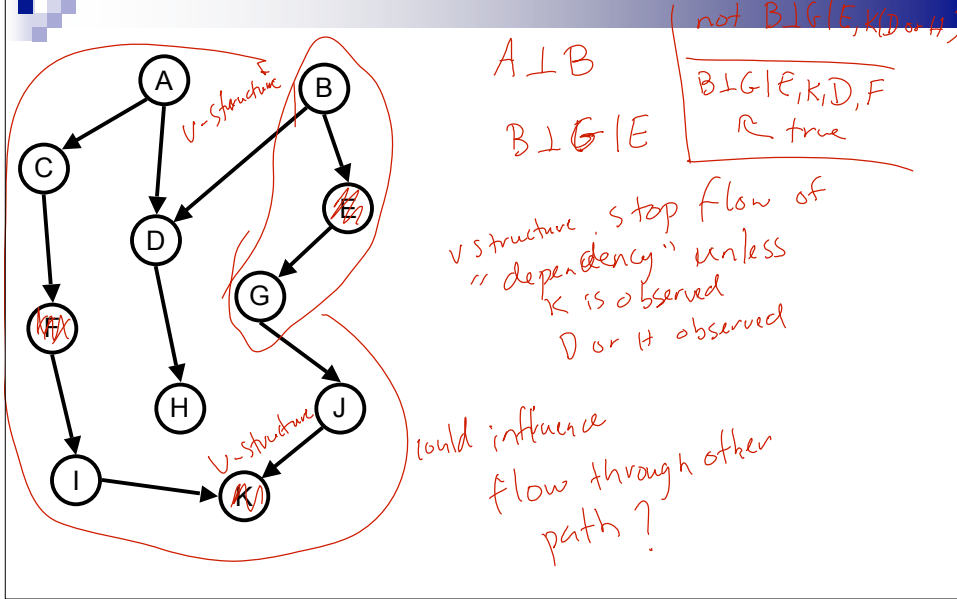


$X \perp Y$

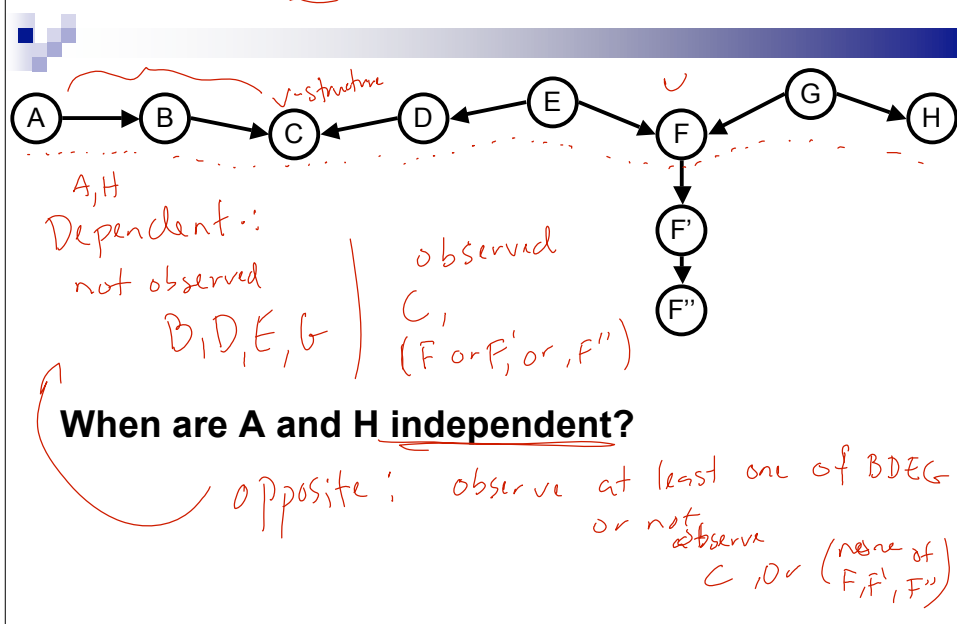
not  $X \perp Y \mid Z$

# Understanding independencies in BNs

## – Some examples



## An active trail – Example



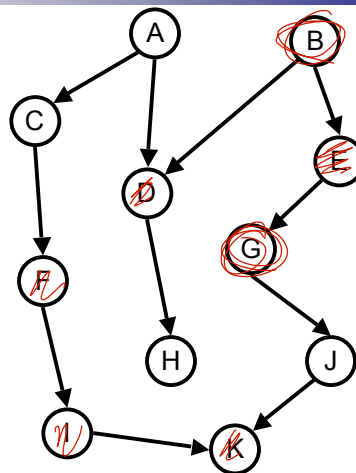
## Active trails formalized

*dependencies can flow*

- A path  $X_1 - X_2 - \dots - X_k$  is an **active trail** when variables  $\mathbf{O} \cap \{X_1, \dots, X_n\}$  are observed if for each consecutive triplet in the trail:
  - $X_{i-1} \rightarrow X_i \rightarrow X_{i+1}$ , and  $X_i$  is **not observed** ( $X_i \notin \mathbf{O}$ )
  - $X_{i-1} \leftarrow X_i \leftarrow X_{i+1}$ , and  $X_i$  is **not observed** ( $X_i \notin \mathbf{O}$ )
  - $X_{i-1} \leftarrow X_i \rightarrow X_{i+1}$ , and  $X_i$  is **not observed** ( $X_i \notin \mathbf{O}$ )
  - $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$ , and  $X_i$  is **observed** ( $X_i \in \mathbf{O}$ ), or **one of its descendants**

## Active trails and independence?

- **Theorem:** Variables  $X_i$  and  $X_j$  are independent given  $Z \cap \{X_1, \dots, X_n\}$  if there is **no active trail** between  $X_i$  and  $X_j$  when variables  $Z \cap \{X_1, \dots, X_n\}$  are observed



# The BN Representation Theorem

If conditional independencies in BN are subset of conditional independencies in  $P$

Obtain

Joint probability distribution:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | \text{Pa}_{X_i})$$

Important because:

Every  $P$  has at least one BN structure  $G$

*some BNs can represent many  $P$  using active trails*

If joint probability distribution:  
 $P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | \text{Pa}_{X_i})$

Obtain

Then conditional independencies in BN are subset of conditional independencies in  $P$

Important because:

Read independencies of  $P$  from BN structure  $G$