Topics in Probabilistic and Statistical Databases

Lecture 8: Implicit Probabilistic Data

Dan Suciu
University of Washington

Equivalences

- Let $A, B \in \text{STRUCT}[\sigma]$

- They are *elementary equivalent,* $A \equiv B$,
  if $\forall\ \phi$, $A \models \phi$ iff $B \models \phi$

- They are *isomorphic,* $A \cong B$,
  if there exists a isomorphism $f : A \rightarrow B$

Equivalences

- If $A \cong B$, then $A \equiv B$ [why ?]
- There are A, B s.t. $A \equiv B$ and not $A \cong B$ [give examples in class]
- If A is finite and $A \equiv B$ then $A \cong B$ [why ?]

Partial Isomorphism

- Let $\underline{c} = (c_1, \ldots, c_n)$ be the constants in σ
- Let A, B be two structures in STRUCT[σ]

Definition A partial isomorphism is given by $\underline{a} = (a_1, \ldots, a_m)$ and $\underline{b} = (b_1, \ldots, b_m)$ s.t. the structures $(\underline{a}, \underline{c})$ and $(\underline{b}, \underline{c})$ are isomorphic.

Ehrenfeucht-Fraisse Games

Given two structures A, B

Two players: **spoiler** and **duplicator** play k rounds

Round i (i =1, …, k):

1. Spoiler picks a structure A or B
2. Spoiler picks an element in that structure $a_i \in A$ or $b_i \in B$
3. Duplicator responds by picking an element in the other structure, $b_i \in B$ or $a_i \in A$

Duplicator wins if $\underline{a}$, $\underline{b}$ is a partial isomorphism

Ehrenfeucht-Fraisse Games

- If duplicator has a winning strategy in k rounds then we write $A \equiv_k B$

- Note: if $A \equiv_n B$ and $k < n$ then $A \equiv_k B$

Quantifier Rank

- The *quantifier rank* of a formula ϕ is defined inductively:

  $qr(t_1 = t_2) = qr(R(t_1,\ldots,t_n)) = 0$
  $qr(\phi_1 \wedge \phi_2) = \max(qr(\phi_1), qr(\phi_2))$
  $qr(\exists\, x.\phi) = 1 + qr(\phi)$
  etc

- FO[k] = all formula ϕ s.t. $qr(\phi) \leq k$

Ehrenfeucht-Fraisse Games

Theorem The following two are equivalent:

- A and B agree on FO[k]
- $A \equiv_k B$

We omit the proof

For now, let's start playing the game !

Games on Sets

- Let $\sigma = \emptyset$ (i.e. no relation names, no constants), and assume $|A| \geq k$, $|B| \geq k$.

Theorem $A \equiv_k B$. [why ?]

Corollary EVEN is not expressible in FO when $\sigma = \emptyset$

Games on Linear Orders

Theorem Let $k > 0$ and L_1, L_2 be linear orders of length $\geq 2^k$. Then $L_1 \equiv_k L_2$.

Proof 1

Define:

- $a_{-1} = \min(L_1)$, $a_0 = \max(L_1)$
- $b_{-1} = \min(L_2)$, $b_0 = \max(L_2)$

Let $\underline{a} = (a_{-1}, a_0, a_1, \ldots, a_i)$, $\underline{b} = (b_{-1}, b_0, b_1, \ldots, b_i)$

Duplicator plays such that $\forall$ j, l:

- If $d(a_j, a_l) < 2^{k-i}$ then $d(a_j, a_l) = d(b_j, b_l)$
- If $d(a_j, a_l) \geq 2^{k-i}$ then $d(b_j, b_l) \geq 2^{k-i}$
- $a_j \leq a_l$ iff $b_j \leq b_l$

Why can the duplicator play like that ?

Proof 2

Remark: if $L_1 \equiv_k L_2$ then the duplicator has a winning strategy where it responds with $\underline{\min}(L_2)$ to $\underline{\min}(L_1)$, and with $\underline{\max}(L_2)$ to $\underline{\max}(L_1)$, and vice versa. [why ?]

Lemma If $L_1^{\leq a} \equiv_k L_2^{\leq b}$ and $L_1^{\geq a} \equiv_k L_2^{\geq b}$ then $(L_1, a) \equiv_k (L_2, b)$

[how does this help us prove the theorem ?]

EVEN

Corollary EVEN is not expressible in FO($<$)

[why ?]

Corollary Finite graph connectivity (CONN) is not expressible in FO

[proof in class]

Random Graphs

- J. Spencer, *The Strange Logic of Random Graphs*
- Binary relation R
  - Classical random graphs: R = undirected
  - We: R = directed (this is standard in databases)

Random Graphs

- Let $n > 0$, and $p = p(n)$ a number in $[0,1]$
  - Examples: $p(n) = ½$, or $p(n) = 1/n^2$ or, $p(n) = 1/n$
- $G(n,p)$ = probability space over graphs with:
  - Nodes = $\{1,2,\ldots,n\}$
  - Each edge has probability $p(n)$
- Main problem: study $\lim_{n\to\infty}\mu_n(A)$

Example

- Let A = “the graph R has a triangle”

- Equivalently: A = $\exists x.\exists y.\exists z.R(x,y),R(y,z),R(z,x)$

- Question: What is $\lim_{n\to\infty}\mu_n(A)$?

Example

- Let A = "the graph R has a triangle"
- Equivalently: A = $\exists x.\exists y.\exists z.R(x,y),R(y,z),R(z,x)$
- Question: What is $\lim_{n\to\infty}\mu_n(A)$?
- Answer: 0 if $p \ll 1/n$, and 1 if $p \gg 1/n$

[Erdos&Reny:1959]

Erdos and Reny's Random Graphs

Now let p = p(n) be a function of n

Theorem [Erdos&Reniy:1959]
For any monotone A, $\exists$ a threshold function t(n) s.t.:

if $p(n) \ll t(n)$ then $\lim_{n\to\infty}\mu_n(A)=0$

if $p(n) \gg t(n)$ then $\lim_{n\to\infty}\mu_n(A)=1$

0/1 Laws

- FO has a 0/1 law on $G(n,p)$ if for any sentence A, $\lim_{n\to\infty}\mu_n(A)=0$ or 1

0/1 Laws

- If G(n,p) has a 0/1 law, denote T its *theory*:
  - $T = \{A \mid \lim_{n\to\infty} \mu_n(A) = 1\}$

- T is a complete theory [WHY ???]
- T has no finite models [WHY ???]
- Goals:
  - Axiomatize T
  - Describe its countable model(s)

Example

- Consider p= ½ , hence G(n, ½).
  - $T = \{A \mid \lim_{n\to\infty} \mu_n(A) = 1\}$
- Q: What is a finite axiomatization of T ?
- Q: What are its countable models ?

Example

- Consider p= ½ , hence G(n, ½).
  - $T = \{A \mid \lim_{n\to\infty} \mu_n(A) = 1\}$
- Q: What is a finite axiomatization of T ?
- A: the extension axioms [Alice's restaurant]
- Q: What are its countable models ?
- A: only one, the Rado graph (random graph)

Threshold Functions v.s. 0/1 Law

- Now consider properties A that are in FO
- Suppose p(n) is such that:
  - For every FO formula A, $\lim_{n \to \infty} \mu_n(A)$ exists
  - For every monotone A, p(n) is not its threshold
- Then: the positive part of FO has 0/1 law
- What to expect:
  - FO admits 0/1 laws *except* at threshold functions

[Erdos&Reny:1959; Spencer:2001]

The Evoluation of Random Graphs

The tuple probability p(n) “grows” from 0 to 1.
How does the random graph evolve ?

$$\boxed{0} \quad \boxed{\frac{1}{n^2}} \quad \frac{1}{n^{3/2}} \quad \frac{1}{n^{4/3}} \quad \frac{1}{n^{5/4}} \quad \ldots \quad \boxed{\frac{1}{n}} \quad \frac{1}{n \cdot ln(n)} \quad \boxed{1}$$

Remark: $C(n) = E[\ Size(R)\ ] \simeq n^2 p(n)$

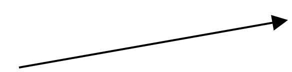
The expected size C(n) “grows” from 0 to n^2.
How does the random graph evolve ?

[Spencer:2001]

The Void

$p(n) \ll 1/n^2$

$C(n) \ll 1$

Contains almost surely	Does not contain almost surely
(nothing)	

The graph is empty

0/1 Law holds

The Void

- Q: what is the theory T ?

- Q: what are the countable models ?

The Void

- Q: what is the theory T ?
  - For every k: "there exists k vertices"
  - $\forall x. \forall y.$ not(R(x,y))
- Q: what are the countable models ?
  - The graph with countable nodes, empty edges
  - Aleph0-categorical !

First Threshold Function

$p(n) = b/n^2$

$C(n) = b$

Q: threshold function for which A ?

No 0/1 Law

First Threshold Function

$p(n) = b/n^2$

$C(n) = b$

Q: threshold function for which A ?

A: $\exists x. \exists y.\ R(x,y)$

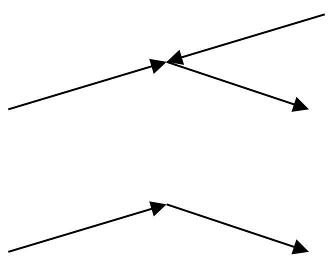
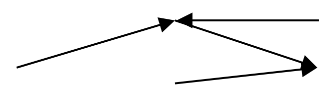
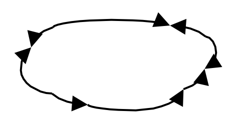
No 0/1 Law

[Spencer:2001]

On the k'th Day

$1/n^{1+1/(k-1)} \ll p(n) \ll 1/n^{1+1/k}$

$n^{1-1/(k-1)} \ll C(n) \ll n^{1-1/k}$

Contains almost surely	**Does not contain almost surely**
trees with $\leq$ k edges 	trees $>$ k edges  cycles 

The graph is disconnected

0/1 Law holds

On the k'th Day

- Q: what is the theory T ?

- Q: what are the countable models ?

On the k'th Day

- Q: what is the theory T ?
  - Every tree with at most k+1 vertices occurs at least r (> 0) times
  - No cycle of size s, for all s > 2
  - No connected components with > k+1 vertices
  - [WRITE ALL THIS IN FO !!]
- Q: what are the countable models ?
  - Infinite copies of each tree with ≤ k+1 vertices
  - Aleph0-categorical !

k’th hreshold Function

$p(n) = b/n^{1+1/k}$

$C(n) = b*n^{1-1/k}$

Q: threshold function for which A ?

No 0/1 Law

k'th hreshold Function

$p(n) = b/n^{1+1/k}$

$C(n) = b*n^{1-1/k}$

Q: threshold function for which A ?

A: $\exists x_0 \ldots \exists x_k.\ R(x_0,x_1),\ldots,R(x_{k-1},x_k)$

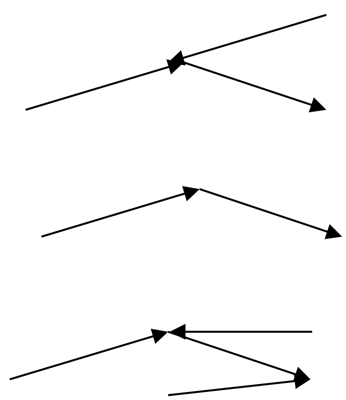
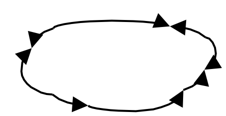
No 0/1 Law

[Spencer:2001]

On Day ω

$1/n^{1+\varepsilon} \ll p(n) \ll 1/n, \quad \forall\, \varepsilon > 0$

$n^{1-\varepsilon} \ll C(n) \ll n, \quad \forall\, \varepsilon > 0$

Contains almost surely	**Does not contain almost surely**
Any Tree 	cycles 

The graph is disconnected

0/1 Law holds

On Day ω

- Q: what is the theory T ?
  - No cycles
  - Any finite tree occurs at least r (> 0) times
- Q: what are the countable models ?
  - Infinite copies of each tree with $\leq$ k+1 vertices
  - May or may not have infinite trees !
  - No longer Aleph0-categorical !
  - But any two models are *elementary equivalent*

ω’th hreshold Function

$p(n) = 1/n$

$C(n) = n$

Q: threshold function for which A ?

No 0/1 Law

ω'th hreshold Function

$p(n) = 1/n$

$C(n) = n$

Q: threshold function for which A ?

A: e.g. triangle: $\exists x.\exists y.\exists z.\ R(x,y),R(y,z),R(z,x)$

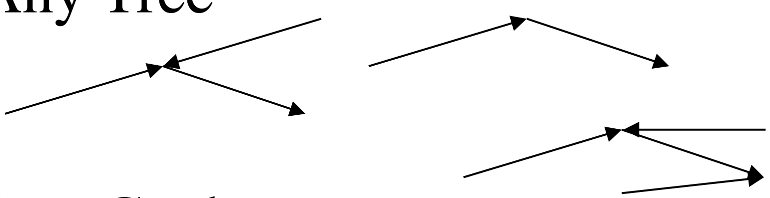
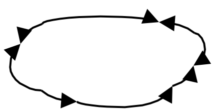
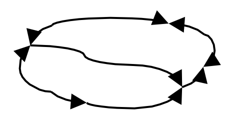
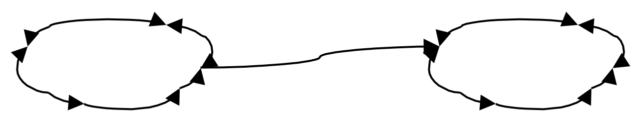
No 0/1 Law

[Spencer:2001]

Past the Double Jump (1/n)

$1/n \ll p(n) \ll \ln(n)/n$

$n \ll C(n) \ll n \ln(n)$

Contains almost surely	**Does not contain almost surely**
Any Tree  Any Cycle 	Any subgraph with k nodes and $\geq$ k+1 edges  

The graph is disconnected

0/1 Law holds

Past the Double Jump (1/n)

- The theory T:
  - Forall k, r: at least r cycles of length k
  - Forall r, finite tree t: at least r copies of t
  - Forall k: no k vertices with k+1 edges
  - Forall k, d, s: not (exists cycle of length k, node x at distance s from the cycle, with degree < d)
- Countable models:
  - Countable copies of each finite tree
  - Countable copies of: cycle k + infinite tree
  - May or may not have infinite trees

The threshold of connectivity

$$p(n) = \ln(n)/n + c/n$$

Theorem [Erdos&Renyi]

$\lim \Pr[G(n,p) \text{ is connected}] = \exp(-\exp(-c))$

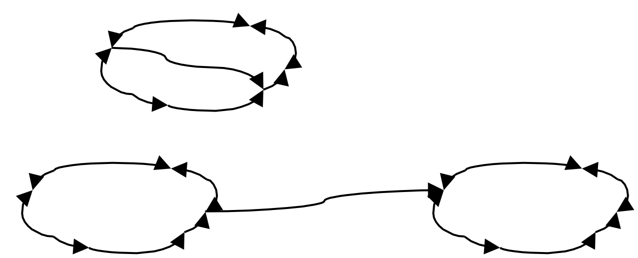
No 0/1 Law

[Spencer:2001]

Past Connectivity

$\ln(n)/n \ll p(n) \ll 1/n^{1-\varepsilon}, \forall\, \varepsilon$

$n \ln(n) \ll C(n) \ll n^{1+\varepsilon}, \forall\, \varepsilon$

Contains almost surely	**Does not contain almost surely**
Every node has degree $\geq$ k, for every k $\geq$ 0 Strange logic of random graphs !!	Any subgraph with k nodes and $\geq$ k+1 edges 

The graph is connected !

0/1 Law holds

Past Connectivity

- The theory T:
  - For every k: there do not exists k veritices with at least k+1 edges
  - All vertices have at least d neighbors
  - For all r, k: there are at least r copies of a cycle of length k

- The countable models:
  - Unicycles followed by infinite trees
  - May or may not have infinite trees

[Spencer:2001]

Big Graphs

$p(n) = 1/n^{\alpha}, \ \alpha \in (0,1)$

$C(n) = n^{2-\alpha}, \alpha \in (0,1)$

α is irrational $\Rightarrow$ 0/1 Law holds

α is rational $\Rightarrow$ 0/1 Law does not hold

Fagin's framework: $\alpha = 0$ 0/1 Law holds

$p(n) = O(1)$

$C(n) = O(n^2)$