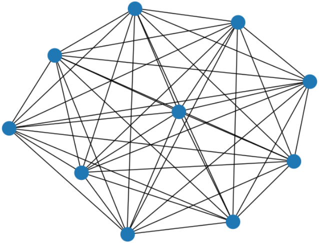
Relating Description Complexity to Entropy

Reijo Jaakkola
reijo.jaakkola@tuni.fi

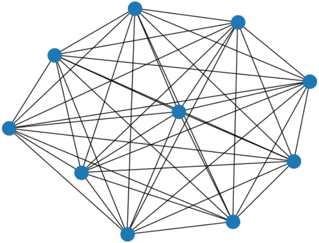
Tampere University

March 16, 2023

What structures are difficult to describe?

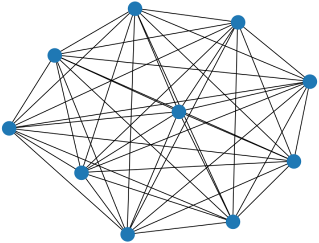


What structures are difficult to describe?

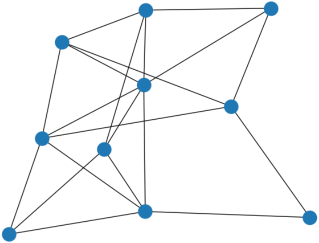


(a) $\forall x \forall y E(x, y)$

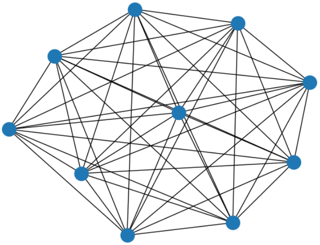
What structures are difficult to describe?



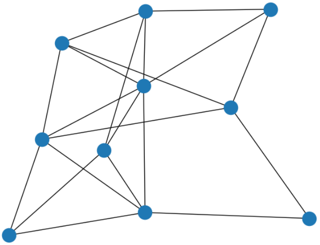
(a) $\forall x \forall y E(x, y)$



What structures are difficult to describe?

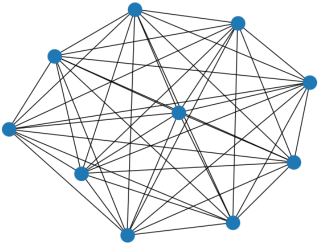


(a) $\forall x \forall y E(x, y)$

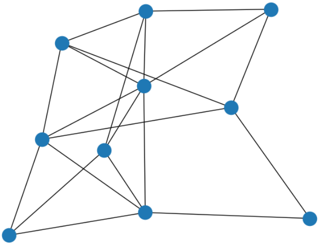


(b) Too difficult...

What structures are difficult to describe?



(a) $\forall x \forall y E(x, y)$



(b) Too difficult...

Intuition: the harder a structure is to describe, the more *random* it is (and vice versa).

Description complexity

Fix a logic $\mathcal{L}$ and a finite collection $\mathcal{M}$ of models.

Description complexity

Fix a logic $\mathcal{L}$ and a finite collection $\mathcal{M}$ of models. Define

$$\mathfrak{M} \equiv \mathfrak{N} \iff \forall \phi \in \mathcal{L} : (\mathfrak{M} \models \phi \iff \mathfrak{N} \models \phi)$$

Given $\mathfrak{M} \in \mathcal{M}$ define $[\mathfrak{M}]_{\equiv} := \{\mathfrak{N} \in \mathcal{M} \mid \mathfrak{N} \equiv \mathfrak{M}\}$.

Description complexity

Fix a logic $\mathcal{L}$ and a finite collection $\mathcal{M}$ of models. Define

$$\mathfrak{M} \equiv \mathfrak{N} \iff \forall \phi \in \mathcal{L} : (\mathfrak{M} \vDash \phi \iff \mathfrak{N} \vDash \phi)$$

Given $\mathfrak{M} \in \mathcal{M}$ define $[\mathfrak{M}]_{\equiv} := \{\mathfrak{N} \in \mathcal{M} \mid \mathfrak{N} \equiv \mathfrak{M}\}$.

Definition

The $\mathcal{L}$-**description complexity** $C_{\mathcal{L}}(M)$ of $M \in \mathcal{M}/\equiv$ is

$$\min\{\text{size}(\phi) : \mathfrak{N} \vDash \phi \text{ iff } \mathfrak{N} \in M\}$$

Description complexity

Fix a logic $\mathcal{L}$ and a finite collection $\mathcal{M}$ of models. Define

$$\mathfrak{M} \equiv \mathfrak{N} \iff \forall \phi \in \mathcal{L} : (\mathfrak{M} \models \phi \iff \mathfrak{N} \models \phi)$$

Given $\mathfrak{M} \in \mathcal{M}$ define $[\mathfrak{M}]_{\equiv} := \{\mathfrak{N} \in \mathcal{M} \mid \mathfrak{N} \equiv \mathfrak{M}\}$.

Definition

The $\mathcal{L}$-**description complexity** $C_{\mathcal{L}}(M)$ of $M \in \mathcal{M}/\equiv$ is

$$\min\{\text{size}(\phi) : \mathfrak{N} \models \phi \text{ iff } \mathfrak{N} \in M\}$$

We are especially interested in the setting where $\mathcal{M}$ consists of all τ-models with domain $\{1, \ldots, n\}$, for some fixed relational vocabulary τ.

Most classes have high description complexity

Fix a signature τ containing at least one relation of arity ≥ 2. Let $m = \max\{ar(R) \mid R \in \tau\}$. Set $\mathcal{L} = \mathrm{FO}[\tau]$ and $\mathcal{M}$ = "τ-models with domain $\{1, \ldots, n\}$". In this case $\equiv$ is just the isomorphism relation.

Most classes have high description complexity

Fix a signature τ containing at least one relation of arity ≥ 2. Let $m = \max\{ar(R) \mid R \in \tau\}$. Set $\mathcal{L} = \mathrm{FO}[\tau]$ and $\mathcal{M}$ = "τ-models with domain $\{1, \ldots, n\}$". In this case $\equiv$ is just the isomorphism relation.

Theorem (J., Kuusisto & Vilander, 2023)

As $n \to \infty$ with high probability we have $C(M) = \Omega(\frac{n^m}{\log(n)})$ for a random isomorphism class M.

Most classes have high description complexity

Fix a signature τ containing at least one relation of arity ≥ 2. Let $m = \max\{ar(R) \mid R \in \tau\}$. Set $\mathcal{L} = \mathrm{FO}[\tau]$ and $\mathcal{M}$ = "τ-models with domain $\{1, \ldots, n\}$". In this case $\equiv$ is just the isomorphism relation.

Theorem (J., Kuusisto & Vilander, 2023)

As $n \to \infty$ with high probability we have $C(M) = \Omega(\frac{n^m}{\log(n)})$ for a random isomorphism class M.

Proof.

The ratio of "short" formulas and isomorphism classes tends to zero as $n \to \infty$. □

Most classes have high description complexity

Fix a signature τ containing at least one relation of arity ≥ 2. Let $m = \max\{ar(R) \mid R \in \tau\}$. Set $\mathcal{L} = \mathrm{FO}[\tau]$ and $\mathcal{M}$ = "τ-models with domain $\{1, \ldots, n\}$". In this case $\equiv$ is just the isomorphism relation.

Theorem (J., Kuusisto & Vilander, 2023)

As $n \to \infty$ with high probability we have $C(M) = \Omega(\frac{n^m}{\log(n)})$ for a random isomorphism class M.

Proof.

The ratio of "short" formulas and isomorphism classes tends to zero as $n \to \infty$. □

Proving such a lower bound for an *explicit* isomorphism class seems to be difficult.

Simple fragment of FO

To prove lower bounds for concrete classes, we consider a weaker logic.

Simple fragment of FO

To prove lower bounds for concrete classes, we consider a weaker logic.

Definition

Fix a propostional vocabulary $\tau = \{p_1, \ldots, p_k\}$. The set of formulas $\mathrm{GMLU}[\tau]$ is defined by the following grammar:

$$\phi ::= p \mid \neg p \mid \phi \vee \phi \mid \phi \wedge \phi \mid \blacklozenge^{\geq d} \phi \mid \blacksquare^{<d} \phi$$

Simple fragment of FO

To prove lower bounds for concrete classes, we consider a weaker logic.

Definition

Fix a propostional vocabulary $\tau = \{p_1, \ldots, p_k\}$. The set of formulas $\mathrm{GMLU}[\tau]$ is defined by the following grammar:

$$\phi ::= p \mid \neg p \mid \phi \vee \phi \mid \phi \wedge \phi \mid \blacklozenge^{\geq d} \phi \mid \blacksquare^{< d} \phi$$

Size of a GMLU formula is defined as follows:

1. $\mathrm{size}(p) = \mathrm{size}(\neg p) = 1$.
2. $\mathrm{size}(\psi \vee \chi) = \mathrm{size}(\psi \wedge \chi) = 1 + \mathrm{size}(\psi) + \mathrm{size}(\chi)$.
3. $\mathrm{size}(\blacklozenge^{\geq d} \psi) = \mathrm{size}(\blacksquare^{< d} \psi) = d + \mathrm{size}(\psi)$.

Positive and negative information have the same cost.

Simple fragment of FO

To prove lower bounds for concrete classes, we consider a weaker logic.

Definition

Fix a propostional vocabulary $\tau = \{p_1, \ldots, p_k\}$. The set of formulas $\mathrm{GMLU}[\tau]$ is defined by the following grammar:

$$\phi ::= p \mid \neg p \mid \phi \vee \phi \mid \phi \wedge \phi \mid \blacklozenge^{\geq d}\phi \mid \blacksquare^{<d}\phi$$

Size of a GMLU formula is defined as follows:

1. $\mathrm{size}(p) = \mathrm{size}(\neg p) = 1$.
2. $\mathrm{size}(\psi \vee \chi) = \mathrm{size}(\psi \wedge \chi) = 1 + \mathrm{size}(\psi) + \mathrm{size}(\chi)$.
3. $\mathrm{size}(\blacklozenge^{\geq d}\psi) = \mathrm{size}(\blacksquare^{<d}\psi) = d + \mathrm{size}(\psi)$.

Positive and negative information have the same cost.

Definition

Given a Kripke model $\mathfrak{M}$ and a world w we define

1. $\mathfrak{M}, w \Vdash \blacklozenge^{\geq d}\psi$ iff there exists $X \subseteq \mathrm{dom}(\mathfrak{M})$ such that $|X| \geq d$ and $\mathfrak{M}, v \Vdash \psi$, for every $v \in X$.
2. $\mathfrak{M}, w \Vdash \blacksquare^{<d}\psi$ iff there exists $X \subseteq \mathrm{dom}(\mathfrak{M})$ such that $|X| < d$ and $\mathfrak{M}, v \Vdash \psi$, for every $v \notin X$.

Lower bounds on description complexity

Given a Kripke model $\mathfrak{M}$ we define

$$\mathfrak{M} \Vdash \phi \text{ iff } \forall w \in \mathrm{dom}(\mathfrak{M}) : \mathfrak{M}, w \Vdash \phi$$

$\mathrm{GMLU}[\tau]$ can define Kripke models of size n up to isomorphism by counting the number of types each 1-type is realized. Thus $\equiv$ is the isomorphism relation.

Lower bounds on description complexity

Given a Kripke model $\mathfrak{M}$ we define

$$\mathfrak{M} \Vdash \phi \text{ iff } \forall w \in \mathrm{dom}(\mathfrak{M}) : \mathfrak{M}, w \Vdash \phi$$

$\mathrm{GMLU}[\tau]$ can define Kripke models of size n up to isomorphism by counting the number of types each 1-type is realized. Thus $\equiv$ is the isomorphism relation.

Theorem (J., Kuusisto & Vilander, 2023)

Let $\mathcal{M}$ be the set of all Kripke models over τ with domain $\{1, \ldots, n\}$.

1. *For every $M \in \mathcal{M}/\equiv$ we have that*

$$C_{\mathrm{GMLU}}(M) \geq \min(n, 2(n-t)),$$

where $t = \max\{t_\pi \mid \pi$ is realized t_π-times in every model in $M\}$.

2. *For every $M \in \mathcal{M}/\equiv$ whose models realize each 1-type sufficiently many times we have that*

$$C_{\mathrm{GMLU}}(M) \geq n + |\tau| 2^{|\tau|+1} - 1.$$

Connection to Boltzmann entropy

The entropy of a "macrostate" is $k_B \ln(|\Omega|)$, where Ω is the number of "microstates" corresponding to the macrostate and k_B is a Boltzmann constant.

Connection to Boltzmann entropy

The entropy of a "macrostate" is $k_B \ln(|\Omega|)$, where Ω is the number of "microstates" corresponding to the macrostate and k_B is a Boltzmann constant.

Intuition: the larger $|\Omega|$ is, the more likely/random the corresponding macrostate is.

Connection to Boltzmann entropy

The entropy of a "macrostate" is $k_B \ln(|\Omega|)$, where Ω is the number of "microstates" corresponding to the macrostate and k_B is a Boltzmann constant.

Intuition: the larger $|\Omega|$ is, the more likely/random the corresponding macrostate is.

Think of microstates as models and macrostates as members of $\mathcal{M}/\equiv$.

Connection to Boltzmann entropy

The entropy of a "macrostate" is $k_B \ln(|\Omega|)$, where Ω is the number of "microstates" corresponding to the macrostate and k_B is a Boltzmann constant.

Intuition: the larger $|\Omega|$ is, the more likely/random the corresponding macrostate is.

Think of microstates as models and macrostates as members of $\mathcal{M}/\equiv$.

Theorem (J., Kuusisto & Vilander, 2023)

Let $\mathcal{M}$ be the set of all Kripke models over τ with domain $\{1, \ldots, n\}$ and $\equiv$ be the isomorphism relation. The largest member of $\mathcal{M}/\equiv$ has maximal GMLU-complexity, namely $n + |\tau|2^{|\tau|+1} - 1$.

Connection to Boltzmann entropy

The entropy of a "macrostate" is $k_B \ln(|\Omega|)$, where Ω is the number of "microstates" corresponding to the macrostate and k_B is a Boltzmann constant.

Intuition: the larger $|\Omega|$ is, the more likely/random the corresponding macrostate is.

Think of microstates as models and macrostates as members of $\mathcal{M}/\equiv$.

Theorem (J., Kuusisto & Vilander, 2023)

Let $\mathcal{M}$ be the set of all Kripke models over τ with domain $\{1, \ldots, n\}$ and $\equiv$ be the isomorphism relation. The largest member of $\mathcal{M}/\equiv$ has maximal GMLU-complexity, namely $n + |\tau| 2^{|\tau|+1} - 1$.

Question

Does a similar result hold in the context of FO?

Connection to expected Boltzmann entropy

$\equiv$ induces a natural probability distribution on $\mathcal{M}$.

$$p_{\equiv}(M) = \frac{|M|}{|\mathcal{M}|}$$

Expected Boltzmann entropy of $\equiv$ is defined as

$$H_B(\equiv) = \sum_{M} p_{\equiv}(M) \log(|M|)$$

Connection to expected Boltzmann entropy

$\equiv$ induces a natural probability distribution on $\mathcal{M}$.

$$p_{\equiv}(M) = \frac{|M|}{|\mathcal{M}|}$$

Expected Boltzmann entropy of $\equiv$ is defined as

$$H_B(\equiv) = \sum_{M} p_{\equiv}(M) \log(|M|)$$

Theorem (J., Kuusisto & Vilander, 2023)

Let $\mathcal{M}$ be the set of all Kripke models over τ with domain $\{1, \ldots, n\}$ and $\equiv$ be the isomorphism relation.

1. $H_B(\equiv) \sim |\tau| n$
2. *Letting* $\langle C_{\mathrm{GMLU}} \rangle := \sum_M p_{\equiv}(M) C_{\mathrm{GMLU}}(M)$ *we have that* $\langle C_{\mathrm{GMLU}} \rangle \sim n$.

In particular $H_B(\equiv) \sim |\tau| \langle C_{\mathrm{GMLU}} \rangle$.

Formula size game for GMLU

The formula size game for $\mathrm{GMLU}[\tau]$ has two players: Samson and Delilah. We refer to them as S and D, or he and she, respectively. The game has three parameters: a natural number $r_0 \geq 1$ and two sets of Kripke-models M_0 and N_0. Positions of the game are of the form (r, M, N) and the starting position is (r_0, M_0, N_0).

Formula size game for GMLU

The formula size game for $\mathrm{GMLU}[\tau]$ has two players: Samson and Delilah. We refer to them as S and D, or he and she, respectively. The game has three parameters: a natural number $r_0 \geq 1$ and two sets of Kripke-models M_0 and N_0. Positions of the game are of the form (r, M, N) and the starting position is (r_0, M_0, N_0).

In each position, S makes a move. The moves available for S in position (r, M, N) are:

Formula size game for GMLU

The formula size game for $\mathrm{GMLU}[\tau]$ has two players: Samson and Delilah. We refer to them as S and D, or he and she, respectively. The game has three parameters: a natural number $r_0 \geq 1$ and two sets of Kripke-models M_0 and N_0. Positions of the game are of the form (r, M, N) and the starting position is (r_0, M_0, N_0).

In each position, S makes a move. The moves available for S in position (r, M, N) are:

- *p*-move: S chooses a τ-literal α. The game ends. If $M \vDash \alpha$ and $N \vDash \neg\alpha$, then S wins. Otherwise D wins.

Formula size game for GMLU

The formula size game for $\mathrm{GMLU}[\tau]$ has two players: Samson and Delilah. We refer to them as S and D, or he and she, respectively. The game has three parameters: a natural number $r_0 \geq 1$ and two sets of Kripke-models M_0 and N_0. Positions of the game are of the form (r, M, N) and the starting position is (r_0, M_0, N_0).

In each position, S makes a move. The moves available for S in position (r, M, N) are:

- *p*-move: S chooses a τ-literal α. The game ends. If $M \vDash \alpha$ and $N \vDash \neg\alpha$, then S wins. Otherwise D wins.
- $\vee$-move: S chooses $M_1, M_2 \subseteq M$ such that $M_1 \cup M_2 = M$ and $r_1, r_2 \geq 1$ such that $r_1 + r_2 + 1 = r$. D chooses whether the next position is (r_1, M_1, N) or (r_2, M_2, N).

Formula size game for GMLU

The formula size game for $\mathrm{GMLU}[\tau]$ has two players: Samson and Delilah. We refer to them as S and D, or he and she, respectively. The game has three parameters: a natural number $r_0 \geq 1$ and two sets of Kripke-models M_0 and N_0. Positions of the game are of the form (r, M, N) and the starting position is (r_0, M_0, N_0).

In each position, S makes a move. The moves available for S in position (r, M, N) are:

- *p*-move: S chooses a τ-literal α. The game ends. If $M \vDash \alpha$ and $N \vDash \neg\alpha$, then S wins. Otherwise D wins.
- $\vee$-move: S chooses $M_1, M_2 \subseteq M$ such that $M_1 \cup M_2 = M$ and $r_1, r_2 \geq 1$ such that $r_1 + r_2 + 1 = r$. D chooses whether the next position is (r_1, M_1, N) or (r_2, M_2, N).
- $\wedge$-move: The same as a $\vee$-move with the roles of A and N switched.

Formula size game for GMLU

The formula size game for $\mathrm{GMLU}[\tau]$ has two players: Samson and Delilah. We refer to them as S and D, or he and she, respectively. The game has three parameters: a natural number $r_0 \geq 1$ and two sets of Kripke-models M_0 and N_0. Positions of the game are of the form (r, M, N) and the starting position is (r_0, M_0, N_0).

In each position, S makes a move. The moves available for S in position (r, M, N) are:

- p-move: S chooses a τ-literal α. The game ends. If $M \vDash \alpha$ and $N \vDash \neg\alpha$, then S wins. Otherwise D wins.
- $\vee$-move: S chooses $M_1, M_2 \subseteq M$ such that $M_1 \cup M_2 = M$ and $r_1, r_2 \geq 1$ such that $r_1 + r_2 + 1 = r$. D chooses whether the next position is (r_1, M_1, N) or (r_2, M_2, N).
- $\wedge$-move: The same as a $\vee$-move with the roles of A and N switched.
- $\blacklozenge^{\geq d}$: S chooses a number $d \in \mathbb{Z}_+$. If $r \leq d$, the game ends and D wins. Otherwise, for every $(\mathfrak{M}, w) \in M$, S chooses d different points $v \in W$. Let M' be the set of models $(\mathfrak{M}, v)$ chosen this way. For every $(\mathfrak{N}, w) \in N$, S chooses $n - (d - 1)$ different points $v \in W$. Let N' be again the set of models chosen. The next position of the game is $(r - d, M', N')$.

Formula size game for GMLU

The formula size game for $\mathrm{GMLU}[\tau]$ has two players: Samson and Delilah. We refer to them as S and D, or he and she, respectively. The game has three parameters: a natural number $r_0 \geq 1$ and two sets of Kripke-models M_0 and N_0. Positions of the game are of the form (r, M, N) and the starting position is (r_0, M_0, N_0).

In each position, S makes a move. The moves available for S in position (r, M, N) are:

- p-move: S chooses a τ-literal α. The game ends. If $M \vDash \alpha$ and $N \vDash \neg\alpha$, then S wins. Otherwise D wins.
- $\vee$-move: S chooses $M_1, M_2 \subseteq M$ such that $M_1 \cup M_2 = M$ and $r_1, r_2 \geq 1$ such that $r_1 + r_2 + 1 = r$. D chooses whether the next position is (r_1, M_1, N) or (r_2, M_2, N).
- $\wedge$-move: The same as a $\vee$-move with the roles of A and N switched.
- $\blacklozenge^{\geq d}$: S chooses a number $d \in \mathbb{Z}_+$. If $r \leq d$, the game ends and D wins. Otherwise, for every $(\mathfrak{M}, w) \in M$, S chooses d different points $v \in W$. Let M' be the set of models $(\mathfrak{M}, v)$ chosen this way. For every $(\mathfrak{N}, w) \in N$, S chooses $n - (d-1)$ different points $v \in W$. Let N' be again the set of models chosen. The next position of the game is $(r - d, M', N')$.
- $\blacksquare^{<d}$: The same as a $\blacklozenge^{\geq d}$ with the roles of M and N switched.

Ongoing & future work

- By bounding the quantifier-depth of GMLU formulas we have been able to obtain more general connections between the size of a model class and its description complexity.

Ongoing & future work

- By bounding the quantifier-depth of GMLU formulas we have been able to obtain more general connections between the size of a model class and its description complexity.
- Extending GMLU with a linear order seems to be a feasible future direction.

Ongoing & future work

- By bounding the quantifier-depth of GMLU formulas we have been able to obtain more general connections between the size of a model class and its description complexity.
- Extending GMLU with a linear order seems to be a feasible future direction.
- It would be natural to consider description complexities of arbitrary model classes, but proving good lower bounds (even for GMLU) in that setting will imply good lower bounds for propositional logic.

Ongoing & future work

- By bounding the quantifier-depth of GMLU formulas we have been able to obtain more general connections between the size of a model class and its description complexity.
- Extending GMLU with a linear order seems to be a feasible future direction.
- It would be natural to consider description complexities of arbitrary model classes, but proving good lower bounds (even for GMLU) in that setting will imply good lower bounds for propositional logic.

Thanks! :–)