Syntactic vs. semantic locality: How good is a cheap approximation?

Chiara Del Vescovo[1] Pavel Klinov[2] Bijan Parsia[1]
Uli Sattler[1] *Thomas Schneider*[3] Dmitry Tsarkov[1]

[1]School of Computer Science, University of Manchester, UK

[2]Institute of Artificial Intelligence, University of Ulm, Germany

[3]Dept. of Computer Science, University of Bremen, Germany

WoMO Graz, 24 July 2012

Overview

1. Syntactic vs. semantic locality

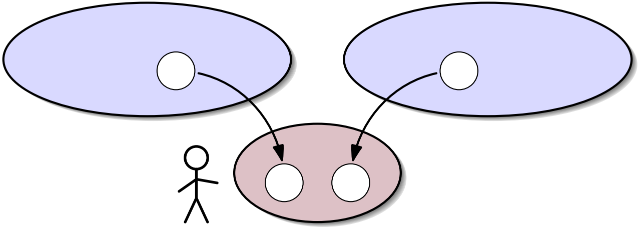
2. How good is a cheap approximation?

And now . . .

1. Syntactic vs. semantic locality
2. How good is a cheap approximation?

A reuse scenario

"Borrow" knowledge from external ontologies

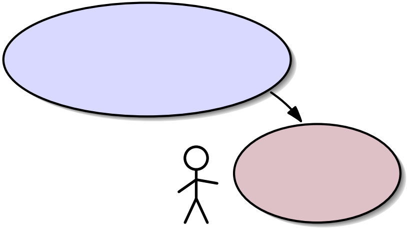


- Provides access to well-established knowledge
- Doesn't require expertise in external disciplines

This scenario is well-understood and implemented.

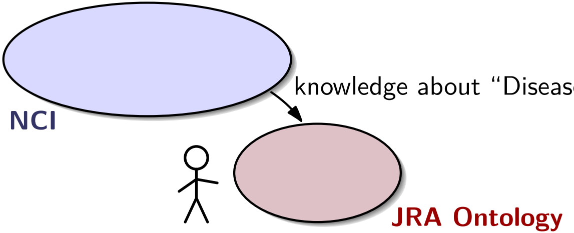
A reuse scenario

Reuse one external, *monolithic* ontology



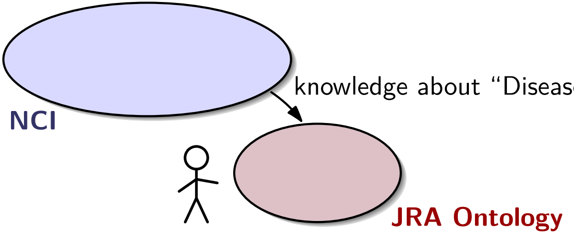
A reuse scenario

Reuse one external, *monolithic* ontology



A reuse scenario

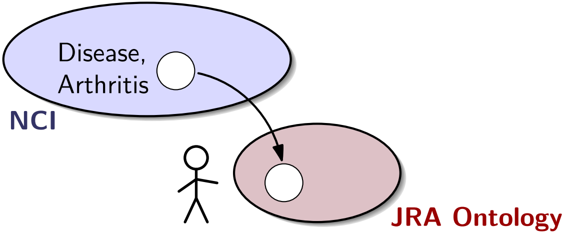
Reuse one external, *monolithic* ontology



How much of **NCI** do we need?

A reuse scenario

Reuse *a part* of an external, monolithic ontology

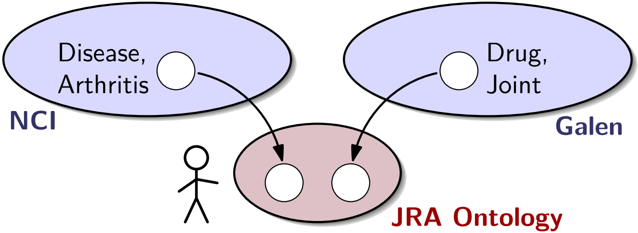


How much of **NCI** do we need?

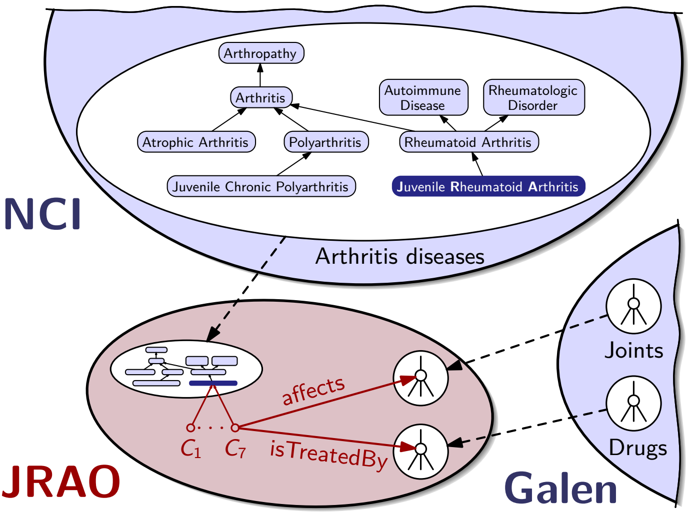
- **Coverage:** Import *everything* relevant for the chosen terms.
- **Economy:** Import *only* what's relevant for them.
  Compute that part quickly.

A reuse scenario

Reuse parts of several external ontologies

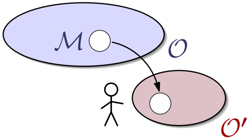


A reuse scenario



What is a module?

$\mathcal{M}$ is a *module* of $\mathcal{O}$ for signature Σ:

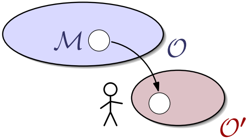


- $\mathcal{M} \subseteq \mathcal{O}$
- $\mathcal{M}$ *covers* $\mathcal{O}$ for Σ, i.e.,

  for all compatible $\mathcal{O}'$,
  $\mathcal{O}' \cup \mathcal{M}$ preserves all knowledge about Σ in $\mathcal{O}' \cup \mathcal{O}$.

What is a module?

$\mathcal{M}$ is a *module* of $\mathcal{O}$ for signature Σ:

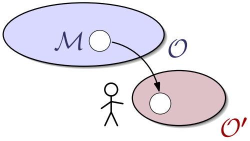


- $\mathcal{M} \subseteq \mathcal{O}$
- $\mathcal{M}$ *covers* $\mathcal{O}$ for Σ, i.e.,

  for all $\mathcal{O}'$ that share only Σ-terms with $\mathcal{O}$,
  for all axioms η built from terms in Σ:

  if η follows from $\mathcal{O}' \cup \mathcal{O}$, then η follows from $\mathcal{O}' \cup \mathcal{M}$.

What is a module?

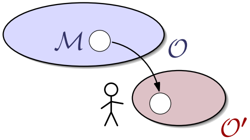


$\mathcal{M}$ is a *module* of $\mathcal{O}$ for signature Σ:

- $\mathcal{M} \subseteq \mathcal{O}$
- $\mathcal{M}$ *covers* $\mathcal{O}$ for Σ, i.e.,

  for all compatible $\mathcal{O}'$,
  $\mathcal{O}' \cup \mathcal{M}$ preserves all knowledge about Σ in $\mathcal{O}' \cup \mathcal{O}$.

What is a module?



$\mathcal{M}$ is a *module* of $\mathcal{O}$ for signature Σ:

- $\mathcal{M} \subseteq \mathcal{O}$
- $\mathcal{M}$ *covers* $\mathcal{O}$ for Σ, i.e.,

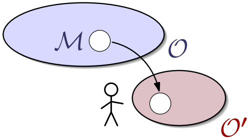
  for all compatible $\mathcal{O}'$,
  $\mathcal{O}' \cup \mathcal{M}$ preserves all knowledge about Σ in $\mathcal{O}' \cup \mathcal{O}$.

- Coverage $\hat{=}$ preserving entailments;

  Without coverage: no encapsulation $\rightsquigarrow$ no *module*

What is a module?

$\mathcal{M}$ is a *module* of $\mathcal{O}$ for signature Σ:



- $\mathcal{M} \subseteq \mathcal{O}$
- $\mathcal{M}$ *covers* $\mathcal{O}$ for Σ, i.e.,

  for all compatible $\mathcal{O}'$,
  $\mathcal{O}' \cup \mathcal{M}$ preserves all knowledge about Σ in $\mathcal{O}' \cup \mathcal{O}$.

- Coverage $\hat{=}$ preserving entailments;
  Without coverage: no encapsulation $\rightsquigarrow$ no *module*

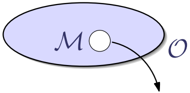
- $\mathcal{O}' \cup \mathcal{O}$ is called *Σ-conservative extension (CE)* of $\mathcal{O}' \cup \mathcal{M}$
  [Ghilardi, Lutz, Wolter 2006]

- Fact: $\mathcal{M}$ covers $\mathcal{O}$ for Σ iff $\mathcal{O}$ is a Σ-CE of $\mathcal{M}$
  $\rightsquigarrow$ $\mathcal{O}'$ doesn't determine what counts as a module

How is a minimal Σ-module extracted?

Simple module extraction algorithm:

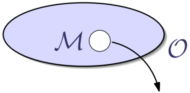
- $\mathcal{M} \leftarrow \mathcal{O}$
- While $\mathcal{M} \setminus \{\alpha\}$ covers $\mathcal{O}$ for Σ,
  for some $\alpha \in \mathcal{M}$,
  remove α from $\mathcal{M}$.
- Output $\mathcal{M}$



How is a minimal Σ-module extracted?

Simple module extraction algorithm:

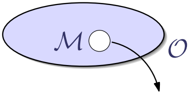
- $\mathcal{M} \leftarrow \mathcal{O}$
- While $\mathcal{M} \setminus \{\alpha\}$ covers $\mathcal{O}$ for Σ,
  for some $\alpha \in \mathcal{M}$,
  remove α from $\mathcal{M}$.
- Output $\mathcal{M}$



Observation:

Different orders of choosing α
can lead to different minimal modules

How is a minimal Σ-module extracted?



Simple module extraction algorithm:

- $\mathcal{M} \leftarrow \mathcal{O}$
- While $\mathcal{M} \setminus \{\alpha\}$ covers $\mathcal{O}$ for Σ,
  for some $\alpha \in \mathcal{M}$,
  remove α from $\mathcal{M}$.
- Output $\mathcal{M}$

Problem:

How to decide the CE property?

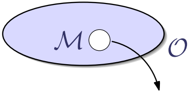
Usually harder than standard reasoning, often undecidable!

[Ghilardi, Lutz, Wolter 2006; Lutz, Walther, Wolter 2007]

Approximation 1: semantic locality ($\emptyset$, Δ)

$\mathcal{M}$ is a $\emptyset$-module of $\mathcal{O}$ for Σ:

if every ax. α in $\mathcal{O} \setminus \mathcal{M}$ is $\emptyset$-local for Σ

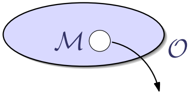


i.e., if all non-Σ symbols are replaced by $\bot$,
then α becomes a tautology [Cuenca Grau et al. 2007]

Approximation 1: semantic locality $(\emptyset, \Delta)$

$\mathcal{M}$ is a $\emptyset$-module of $\mathcal{O}$ for Σ:

if every ax. α in $\mathcal{O} \setminus \mathcal{M}$ is $\emptyset$-local for Σ



i.e., if all non-Σ symbols are replaced by $\bot$,
then α becomes a tautology [Cuenca Grau et al. 2007]

Facts:

- $\emptyset$-module of $\mathcal{O}$ for Σ is uniquely determined
- *the* $\emptyset$-mod$(\Sigma, \mathcal{O})$ covers $\mathcal{O}$ for Σ

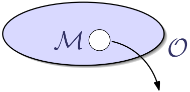
  (but $\mathcal{M}$ isn't necessarily a *minimal* module)

- Deciding semantic locality is as hard as reasoning

Approximation 1: semantic locality ($\emptyset$, Δ)

$\mathcal{M}$ is a $\emptyset$-module of $\mathcal{O}$ for Σ:

if every ax. α in $\mathcal{O} \setminus \mathcal{M}$ is $\emptyset$-local for Σ



i.e., if all non-Σ symbols are replaced by $\bot$,
then α becomes a tautology [Cuenca Grau et al. 2007]

Facts:

- $\emptyset$-module of $\mathcal{O}$ for Σ is uniquely determined
- *the* $\emptyset$-mod($\Sigma, \mathcal{O}$) covers $\mathcal{O}$ for Σ

  (but $\mathcal{M}$ isn't necessarily a *minimal* module)

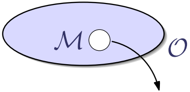
- Deciding semantic locality is as hard as reasoning

Dual notion: Δ-module, Δ-locality

Approximation 2: syntactic locality ($\bot$, $\top$)

$\mathcal{M}$ is a $\bot$-module of $\mathcal{O}$ for Σ:

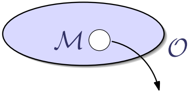
if every ax. α in $\mathcal{O} \setminus \mathcal{M}$ is $\bot$-local for Σ



i.e., α is generated by a grammar
that describes obviously $\emptyset$-local axioms
for the DL $\mathcal{SROIQ}$ underlying OWL 2 [Cuenca Grau et al. 2007]

Approximation 2: syntactic locality ($\bot, \top$)

$\mathcal{M}$ is a $\bot$-module of $\mathcal{O}$ for Σ:



if every ax. α in $\mathcal{O} \setminus \mathcal{M}$ is $\bot$-local for Σ

i.e., α is generated by a grammar
that describes obviously $\emptyset$-local axioms
for the DL $\mathcal{SROIQ}$ underlying OWL 2 [Cuenca Grau et al. 2007]

Facts:

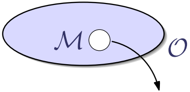
- *The* $\bot$-mod$(\Sigma, \mathcal{O})$ contains the $\emptyset$-mod$(\Sigma, \mathcal{O})$
  and hence covers $\mathcal{O}$ for Σ

  (but again isn't necessarily a *minimal* module)

- syntactic locality can be decided efficiently: in poly-time!

Approximation 2: syntactic locality ($\bot, \top$)

$\mathcal{M}$ is a $\bot$-module of $\mathcal{O}$ for Σ:



if every ax. α in $\mathcal{O} \setminus \mathcal{M}$ is $\bot$-local for Σ

i.e., α is generated by a grammar
that describes obviously $\emptyset$-local axioms
for the DL $\mathcal{SROIQ}$ underlying OWL 2 [Cuenca Grau et al. 2007]

Facts:

- *The* $\bot$-mod($\Sigma, \mathcal{O}$) contains the $\emptyset$-mod($\Sigma, \mathcal{O}$) and hence covers $\mathcal{O}$ for Σ

  (but again isn't necessarily a *minimal* module)

- syntactic locality can be decided efficiently: in poly-time!

Dual notion: $\top$-module, $\top$-locality

Summary: locality-based modules (LBMs)

- Syntactic LBMs are cheap for DLs up to OWL
- Semantic LBMs are expensive for expressive DLs

  (and infeasible for FOL)
- All LBMs provide coverage, but do not guarantee minimality
- Conservativity-based modules are infeasible for expressive DLs and FOL

And now . . .

1. Syntactic vs. semantic locality
2. How good is a cheap approximation?

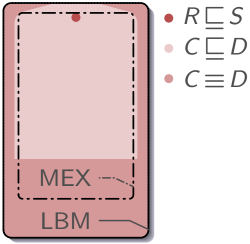
Facts about *syntactic* locality based modules (LBMs)

- $\bot$-mod and $\top$-mod have been implemented: OWL API etc.
- More economic: $\top\bot^*$-mod (alternative nesting until fixpoint)
- Previous experiments: $\top\bot^*$-mod often well-sized in practice
  - Experiments with SNOMED ($\mathcal{EL}$, 350,000 axioms)
  - Compared modules for 24,000 terms from intensive care unit
  - $\top\bot^*$-mod (LBM) $\Leftrightarrow$ module based on model-CE (MEX)

Facts about *syntactic* locality based modules (LBMs)

- $\bot$-mod and $\top$-mod have been implemented: OWL API etc.
- More economic: $\top\bot^*$-mod (alternative nesting until fixpoint)
- Previous experiments: $\top\bot^*$-mod often well-sized in practice
  - Experiments with SNOMED ($\mathcal{EL}$, 350,000 axioms)
  - Compared modules for 24,000 terms from intensive care unit
  - $\top\bot^*$-mod (LBM) $\Leftrightarrow$ module based on model-CE (MEX)
  - Results:

	# axioms	
	MEX	LBM
	10%	15%
	4–5 s	4–7 s



Can you take a little off, please?

Semantic LBMs ($\emptyset$-mod and Δ-mod)

- ... are contained in the respective syntactic LBM, remember:

$$\emptyset\text{-mod}(\Sigma, \mathcal{O}) \subseteq \bot\text{-mod}(\Sigma, \mathcal{O})$$
$$\Delta\text{-mod}(\Sigma, \mathcal{O}) \subseteq \top\text{-mod}(\Sigma, \mathcal{O})$$

- ... are extracted using reasoning
- ... have not been implemented yet

$\rightsquigarrow$ Are they actually (typically, significantly, ...) smaller?
How much more expensive is their extraction?

Questions

Given a signature Σ and ontology $\mathcal{O}$,

1. … how likely is $\emptyset$-mod$(\Sigma, \mathcal{O}) \subset \bot$-mod$(\Sigma, \mathcal{O})$, and how large is the difference?

   (variation: given axiom α,
   is it likely that α is $\emptyset$-local but not $\bot$-local for Σ?)

2. … what is the difference in extraction time?

Questions

Given a signature Σ and ontology $\mathcal{O}$,

1. ... how likely is $\emptyset$-mod$(\Sigma, \mathcal{O}) \subset \bot$-mod$(\Sigma, \mathcal{O})$, and how large is the difference?

   (variation: given axiom α,
   is it likely that α is $\emptyset$-local but not $\bot$-local for Σ?)

2. ... what is the difference in extraction time?

Later: the same questions for the pairs

- Δ-mod vs. $\top$-mod
- $\Delta\emptyset^*$-mod vs. $\top\bot^*$-mod

Sampling the seed signatures

- $\mathcal{O}$ has exponentially many potential seed signatures Σ.
- Modules for different Σ_1, Σ_2 may coincide.
- Still, $\mathcal{O}$ can have exp. many modules. [Del Vescovo et al., 2010]
- We don't yet know what typical seed signatures are.

Sampling the seed signatures

- $\mathcal{O}$ has exponentially many potential seed signatures Σ.
- Modules for different Σ_1, Σ_2 may coincide.
- Still, $\mathcal{O}$ can have exp. many modules. [Del Vescovo et al., 2010]
- We don't yet know what typical seed signatures are.

$\leadsto$ Sample random seed signatures!

  - Sample one Σ: pick each axiom with probability $p = 0.5$
  - Achieve confidence interval $\pm 5\%$ with confidence level 95%: select 400 random Σ's (if $\mathcal{O}$ is big enough)

Sampling the seed signatures

- $\mathcal{O}$ has exponentially many potential seed signatures Σ.
- Modules for different Σ_1, Σ_2 may coincide.
- Still, $\mathcal{O}$ can have exp. many modules. [Del Vescovo et al., 2010]
- We don't yet know what typical seed signatures are.

$\rightsquigarrow$ Sample random seed signatures!

- Sample one Σ: pick each axiom with probability $p = 0.5$
- Achieve confidence interval $\pm 5\%$ with confidence level 95%: select 400 random Σ's (if $\mathcal{O}$ is big enough)

- Non-random seed signatures

  Genuine mod.s (GMs)

  - $\bot$-mod(Sig(α), $\mathcal{O}$), for $\alpha \in \mathcal{O}$
  - every module of $\mathcal{O}$ is the union of some GMs

$\rightsquigarrow$ include axiom signatures Sig(α)

The ontology corpus

Name	Expressivity	#Axioms	Sig. size
BioPortal (148 entries)	$\mathcal{ALCN}$–$\mathcal{SHIN}(\mathcal{D})$/ $\mathcal{SOIN}(\mathcal{D})$	38–4,735	21–3,161
TONES			
Galen	$\mathcal{ALEHIF}+$	4,735	3,161
Koala	$\mathcal{ALCON}(\mathcal{D})$	42	32
Mereology	$\mathcal{SHIN}$	38	21
MiniTambis-rep'd	$\mathcal{ALCN}$	170	227
OWL-S Profile	$\mathcal{ALCHOIN}(\mathcal{D})$	276	163
People	$\mathcal{ALCHOIN}$	108	96
Tambis-full	$\mathcal{SHIN}(\mathcal{D})$	592	497
University	$\mathcal{SOIN}(\mathcal{D})$	52	44

Results I: cheap is good!

1. For 151 out of 156 ontologies, $\emptyset$-mod and $\bot$-mod agree, i.e.:
   - Given an *arbitrary* Σ, there is *no* difference between
     - $\emptyset$-mod($\Sigma, \mathcal{O}$) and $\bot$-mod($\Sigma, \mathcal{O}$), and
     - any α being $\emptyset$-local and $\bot$-local w.r.t. Σ,

     at a significance level of 0.05.
   - Given *any* axiom signature Sig(α), there is *no* difference between $\emptyset$-mod(Sig(α), $\mathcal{O}$) and $\bot$-mod(Sig(α), $\mathcal{O}$).
2. Extracting a $\emptyset$-module took up to $6\times$ as long as $\bot$-module (average $2.7\times$)

Results II: cheap seems good enough

1. For 2 ontologies from BioPortal,[1] negligible differences:
   - Up to 30 out of 3,446 (resp. 6,008) axioms
   - Axioms are: $\mathtt{r} \equiv (\mathtt{r}^-)^-$, for some role (object property) $\mathtt{r}$

     i.e., `EquivObjProps(r, inv(inv(r)))`
   - Uncritical: these are few tautologies

     (Published version of some BioPortal ontologies
     is closed under certain entailments)
2. Extraction time up to $6\times$ on average

[1] Experimental Factor Ontology and Software Ontology

Results III: a single type of culprit

For the remaining 3 ont.s,[2] small differences of 1 common pattern

Example axiom α:

$$\mathtt{M} \;\equiv\; \underline{\mathtt{S}} \sqcap \forall \underline{\mathtt{c}}.\mathtt{F} \sqcap \forall \underline{\mathtt{g}}.\{\mathtt{m}\} \sqcap {=}3\, \underline{\mathtt{c}}.\top$$

```
EquivClasses(M,
S  and  c only F  and  g value m  and  c exactly 3 Thing)
```

[2]Koala, miniTambis and Tambis

Results III: a single type of culprit

For the remaining 3 ont.s,[2] small differences of 1 common pattern

Example axiom α:

$$\mathtt{M} \equiv \underline{\mathtt{S}} \sqcap \forall \underline{\mathtt{c}}.\mathtt{F} \sqcap \forall \underline{\mathtt{g}}.\{\mathtt{m}\} \sqcap {=}3\, \underline{\mathtt{c}}.\top$$

```
EquivClasses(M,
S and c only F and g value m and c exactly 3 Thing)
```

- Suppose $\Sigma = \{\mathtt{S}, \mathtt{c}, \mathtt{g}\}$
- α is not $\bot$-local because none of its conjuncts is $\bot$-equiv.
- α is $\emptyset$-local:
  after replacing $\mathtt{M}, \mathtt{F}$ with $\bot$, it becomes a tautology
  in particular, $\forall \mathtt{c}.\bot \sqcap {=}3\, \mathtt{c}.\top$ cannot have any instances

[2] Koala, miniTambis and Tambis

Results IV: cheap still seems good enough

1. These culprits have
   - no effects on Koala modules

     (only singleton differences for locality)
   - small effects on miniTambis:
     - $\bot$-modules up to 4 axioms (3%) larger than $\emptyset$-modules
     - $\bot$-GMs up to 7 axioms (75%) larger than $\emptyset$-GMs
   - small effects on Tambis:
     - $\bot$-modules up to 11 axioms (2%) larger than $\emptyset$-modules
     - $\bot$-GMs up to 41 axioms (26%) larger than $\emptyset$-GMs
2. Extraction time up to $5\times$ on average for Tambis
   (not measurable for Koala and miniTambis)

Results V: beyond $\emptyset$- vs. $\bot$-locality

Δ-modules cannot always be extracted using DL reasoners:

- Remember – locality check: replace non-Σ symbols with $\top$ and test for tautology
- Global restrictions of $\mathcal{SROIQ}$ don't allow $\top$-role in number restrictions or role chains
- This affects 39 ontologies in our corpus

For the remaining 117 ontologies,
there is no (statistically significant) difference:

- between Δ- and $\top$-modules
- between $\Delta\emptyset^*$- and $\top\bot^*$-modules

Lessons learnt

- No or little difference btn. semantic and syntactic locality

$\rightsquigarrow$ Syntactic locality seems a good approximation of semantic locality

$\rightsquigarrow$ Cheap is good!

- (Still, semantic module extraction often fast in practice)

Outlook

- Incorporate the missing 39 (richer) ontologies into
  - Δ- vs. $\top$-locality
  - $\Delta\emptyset^*$- vs. $\top\bot^*$-locality
- Extend study to larger ontologies
  - NCI has axioms that nest the culprit pattern
  - Not reproducible with the official releases
- Modify sampling
  - Put more weight on small and large seed signatures
  - Measure difference w.r.t. a given *module*
    $\rightsquigarrow$ sampling of modules instead of seed signatures
- Include conservativity-based modules (for lightweight DLs)

Outlook

- Incorporate the missing 39 (richer) ontologies into
  - Δ- vs. $\top$-locality
  - $\Delta\emptyset^*$- vs. $\top\bot^*$-locality
- Extend study to larger ontologies
  - NCI has axioms that nest the culprit pattern
  - Not reproducible with the official releases
- Modify sampling
  - Put more weight on small and large seed signatures
  - Measure difference w.r.t. a given *module*
    $\leadsto$ sampling of modules instead of seed signatures
- Include conservativity-based modules (for lightweight DLs)

Thank you.