Finding Needles in a Haystack: Detecting Misaligned and Missing Concepts in SNOMED CT using Graph and Lexical Patterns

GQ Zhang

Institute for Biomedical Informatics
University of Kentucky
Lexington, Kentucky, USA

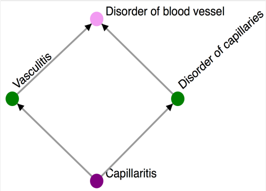
joint work with
Licong Cui, Shiqiang Tao, Wei Zhu, University of Kentucky, USA
Olivier Bodenreider, James Case, National Library of Medicine, USA

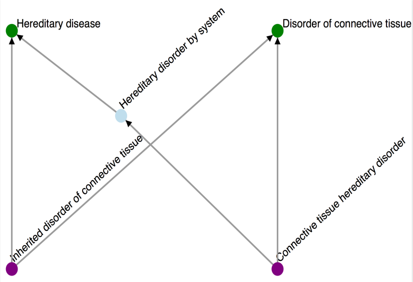
Institute for
Biomedical Informatics

Graphlets

Given a background graph *G*, such as the one represented by the subsumption (is-a) hierarchy of SNOMED CT, a *graphlet g* is a subgraph of *G* such that

- *g* is an induced subgraph of *G*;
- *g* is a "convex" subgraph of *G* generated by a pair of nodes;
- *g* is "small" (like "needles" in a haystack).




Hereditary disease
Hereditary disorder by system
Disorder of connective tissue
Inherited disorder of connective tissue
Connective tissue hereditary disorder

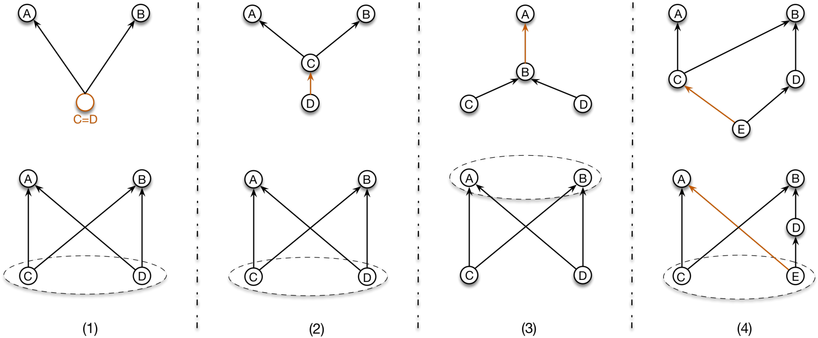
Graphlets from the 09/2014 version

Why graphlets?

- Graph structure provides a natural context for concepts in ontologies as it positions a concept in relationship with other concepts;
- Small graphs are amendable for visualization and interactive exploration, leveraging the power of human visual perception;
- Potential to use graphlets for not only detecting errors, but also coming up with corrections.

What graphlets? (Intuition)

- A lattice is a specific type of directed acyclic graph (DAG) such that any two nodes have a unique maximal common descendant, as well as a unique minimal common ancestor. A lattice is a desirable structural property for a well-formed ontology;
- Errors often lead to abnormal graph patterns – non-lattice graphlets.



What graphlets? (Mathematical theory)

Attributes

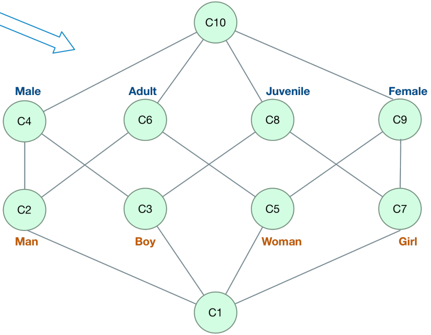
Objects

	Female	Juvenile	Adult	Male
Girl	x	x		
Woman	x		x	
Boy		x		x
Man			x	x

List of Formal Concepts:

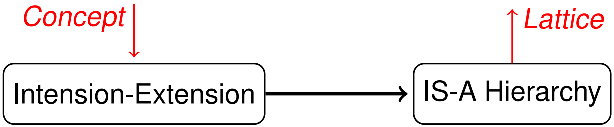
C1	{}, {Female, Juvenile, Adult, Male}
C2	{Man}, {Male, Adult}
C3	{Boy}, {Male, Juvenile}
C4	{Boy, Man}, {Male}
C5	{Woman}, {Female, Adult}
C6	{Woman, Man}, {Adult}
C7	{Girl}, {Female, Juvenile}
C8	{Girl, Boy}, {Juvenile}
C9	{Girl, Woman}, {Female}
C10	{Girl, Woman, Boy, Man}, {}

Extent Intent



Lattice

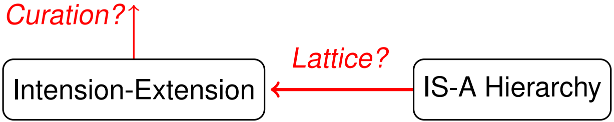
What graphlets – well-structured ontologies are lattices



- Every concept can be looked at from two aspects: intension and extension: intension – intrinsic or endogenous; extension: extrinsic or exogenous;
- For each concept, intension and extension determine each other;
- Conceptual taxonomy (is-a) is a partially ordered set (poset);
- This poset must be a lattice (Formal Concept Analysis).

Z-B: Large-scale, Exhaustive Lattice-based Structural Auditing of SNOMED CT. AMIA Annu Symp Proc. 2010;2010:922-6. PMC3041382.

What graphlets – rationale for lattice-based auditing

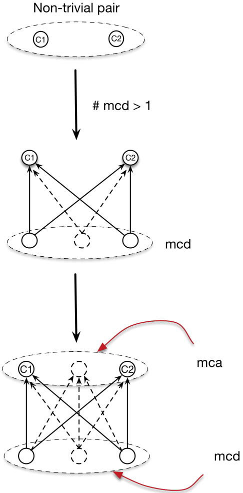


- The is-a hierarchical order is given in SNOMED CT;
- Intension and extension are unspecified and implicit, whose existence is accepted;
- If the taxonomic order is *not* a lattice, then there is a potential *problem*;
- Identifying non-lattice graphlets provides an error-agnostic methodology for “finding needles in a haystack.”

Extracting non-lattice graphlets in SNOMED CT

Material: September 2015 version of SNOMED CT (U.S. edition)

1. Identify non-lattice pairs. $p = (c_1, c_2)$ such that the size of its maximal common descendants $mcd(p)$ is > 1;
2. Construct non-lattice graphlets. Given non-lattice pair $p = (c_1, c_2)$ and its $mcd(p)$,
   - Reversely computing the minimal common ancestors of the maximal common descendants – $mca(mcd(p))$;
   - Aggregating all the concepts and edges between (including) any concept in $mca(mcd(p))$ and any of the maximal common descendants $mcd(p)$. Call $mca(mcd(p))$ the upper boundary, and $mcd(p)$ the lower boundary.



Results (I)

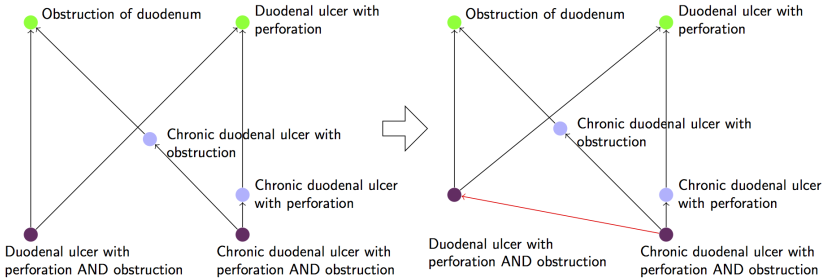
- 631,006 non-lattice pairs were found in the September 2015 version of the SNOMED CT (U.S. edition);
- 171,011 non-lattice graphlets extracted;
- Sizes of non-lattice graphlets ranged from 4 to 5,137;
- 90% of the non-lattice graphlets had sizes 4 to 100;
-
    - 3,339 graphlets of size 4 contained in 28,292 larger graphlets;
    - 3,773 graphlets of size 5 contained in 34,808 larger graphlets;
    - 5,342 graphlets of size 6 contained in 40,404 larger graphlets.
- Useful strategy to focus on small-sized graphlets for auditing work.

Mining SNOMED CT’s non-lattice graphlets: four lexical patterns

- Containment
- Intersection
- Union
- Union-intersection

Containment

- The bag of words for one concept in the upper boundary is contained in the bag of words for another concept in the upper bounary; or
- The bag of words for one concept in the lower boundary is contained in the bag of words for another concept in the lower boundary.



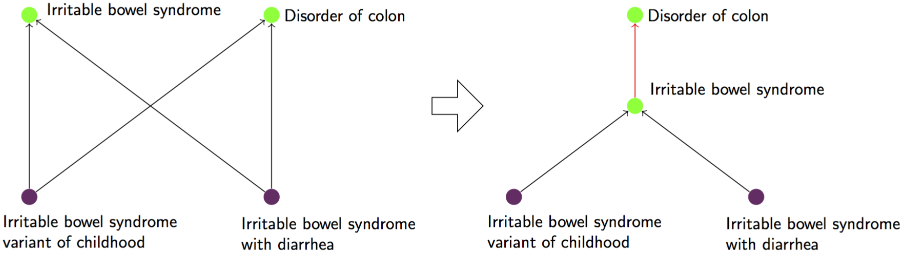
Obstruction of duodenum
Duodenal ulcer with perforation
Chronic duodenal ulcer with obstruction
Chronic duodenal ulcer with perforation
Duodenal ulcer with perforation AND obstruction
Chronic duodenal ulcer with perforation AND obstruction
Obstruction of duodenum
Duodenal ulcer with perforation
Chronic duodenal ulcer with obstruction
Chronic duodenal ulcer with perforation
Duodenal ulcer with perforation AND obstruction
Chronic duodenal ulcer with perforation AND obstruction

$\{\mathit{duodenal}, \mathit{ulcer}, \mathit{perforation}, \mathit{obstruction}\} \subset \{\mathit{chronic}, \mathit{duodenal}, \mathit{ulcer}, \mathit{perforation}, \mathit{obstruction}\}$

This situation generally suggests a missing hierarchical relation between concepts in the upper boundary or the lower boundary.

Intersection

The intersection of bags of words for concepts in the lower boundary is equal to the bag of words for some concept in the upper boundary.


Irritable bowel syndrome
Disorder of colon
Irritable bowel syndrome variant of childhood
Irritable bowel syndrome with diarrhea
Disorder of colon
Irritable bowel syndrome
Irritable bowel syndrome variant of childhood
Irritable bowel syndrome with diarrhea

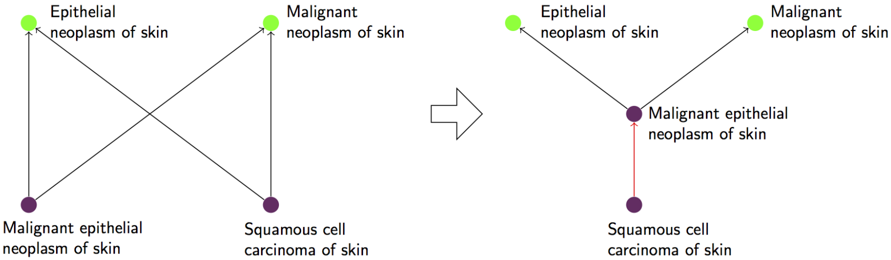
$\{\textit{irritable}, \textit{bowel}, \textit{syndrome}, \textit{variant}, \textit{childhood}\} \cap \{\textit{irritable}, \textit{bowel}, \textit{syndrome}, \textit{diarrhea}\}$

$= \{\textit{irritable}, \textit{bowel}, \textit{syndrome}\}$

This situation generally suggests a missing hierarchical relation between concepts in the upper boundary.

Union

The union of the bags of words for concepts in the upper boundary is equal to the bag of words for some concept in the lower boundary.


Epithelial neoplasm of skin
Malignant neoplasm of skin
Malignant epithelial neoplasm of skin
Squamous cell carcinoma of skin
Epithelial neoplasm of skin
Malignant neoplasm of skin
Malignant epithelial neoplasm of skin
Squamous cell carcinoma of skin

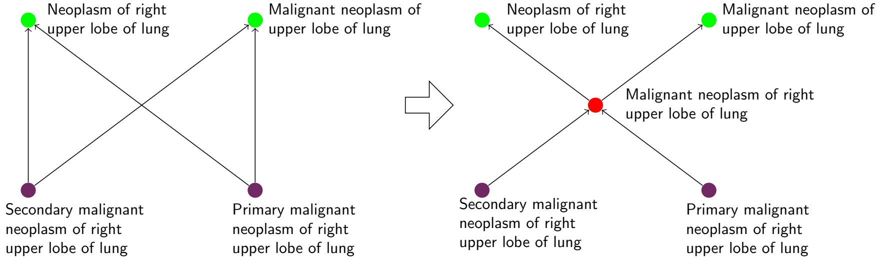
$$\{\textit{epithelial}, \textit{neoplasm}, \textit{skin}\} \cup \{\textit{malignant}, \textit{neoplasm}, \textit{skin}\}$$

$$= \{\textit{epithelial}, \textit{neoplasm}, \textit{skin}, \textit{malignant}\}$$

This situation generally suggests a missing hierarchical relation between concepts in the lower boundary.

Union-intersection (UI)

The union of the bags of words for concepts in the upper boundary is equal to the intersection of bags of words for concepts in the lower boundary.


Neoplasm of right upper lobe of lung
Malignant neoplasm of upper lobe of lung
Secondary malignant neoplasm of right upper lobe of lung
Primary malignant neoplasm of right upper lobe of lung
Neoplasm of right upper lobe of lung
Malignant neoplasm of upper lobe of lung
Malignant neoplasm of right upper lobe of lung
Secondary malignant neoplasm of right upper lobe of lung
Primary malignant neoplasm of right upper lobe of lung

$$\{\mathit{neoplasm}, \mathit{right}, \mathit{upper}, \mathit{lobe}, \mathit{lung}\} \cup \{\mathit{malignant}, \mathit{neoplasm}, \mathit{upper}, \mathit{lobe}, \mathit{lung}\} =$$
$$\{\mathit{secondary}, \mathit{malignant}, \mathit{neoplasm}, \mathit{right}, \mathit{upper}, \mathit{lobe}, \mathit{lung}\} \cap \{\mathit{primary}, \mathit{malignant}, \mathit{neoplasm}, \mathit{right}, \mathit{upper}, \mathit{lobe}, \mathit{lung}\}$$

This situation generally suggests a missing intermediary concept between the upper boundary and the lower boundary.

Results: pattern distribution

	Containment	Intersection	Union	UI	Total
Size 4	160	336	31	17	544
Size 5	229	291	75	13	608
Size 6	347	458	58	31	894
Total	736	1,085	164	61	2,046

Evaluation

To assess the effectiveness of our method in identifying real errors in SNOMED CT, we focused on small non-lattice graphlets following any of the four lexical patterns.

- A random sample of 100 graphlets of sizes 4-6 was selected from the two largest subhierarchies: Clinical finding and Procedure;
- The sample graphlets were rendered in Scalable Vector Graphics (SVGs) to facilitate visualization and evaluation;
- 65 from Clinical Finding; 35 from Procedure;
- 37 – Containment, 46 – Intersection, 13 – Union, 4 – UI;
- Triaged the 100 sample to eliminate most complex cases;
- 59 retained for review by domain experts. Experts confirmed the existence of errors and remediations;
- Error rate $\geq$ 59%, since some erroneous graphlets may not have been selected for review during the triage process.

Discussion

- *Significance*. Our approach not only uncovered novel errors (e.g. the assertion that "A is-a B" is false), but also suggested remediations (*if A is not B, then what A should be?*). Other methods do not address.
- *Graphlets*. Focusing on non-lattice graphlets of smaller size provides an effective way of auditing hierarchical relations in SNOMED CT. Small graphlets are easier to review. Fixing errors in small graphlets may mechanically fixing those in larger graphlets.
- *Practical quality impact*. Impact includes value set definition for EHR decision support, quality reporting and cohort selection. Some errors involve concepts from the CORE Problem List.
- *Generalization*. Virtually all biomedical ontologies are organized into subsumption hierarchies. Our approach is generalizable.
- *Related work*. Our work is very different from abstraction networks (AbNs): AbNs nodes are groups of concepts; AbNs rely on outgoing attribute relationships for grouping concepts into areas.

Limitations

- Our suggested remediation is based on the inferred concept hierarchy. A more meaningful remediation would be to modify the logical definitions, so that the appropriate hierarchy can be inferred.
- We only reported the lower bound of the rate of identified errors. Our choice was justified by (1) the need to minimize the workload of medical experts in this labor-intensive review process; (2) the purpose of the evaluation was to show the promise of combining non-lattice graphlets and lexical patterns to not only detect potential errors in SNOMED CT, but also facilitate remediation (as a proof of principle).
- Not all non-lattice graphlets fall into the four lexical patterns.
- The remediation suggested by the presence of a lexical pattern is not always accurate.
- Non-lattice graphlets may reveal modeling problems in SNOMED CT, but they may not be easily fixed by adding a missing is-a relation or a missing concept.

Conclusion

- We introduced a novel hybrid approach using non-lattice graphlets and lexical information in concept names for detecting missing hierarchical relations or missing concepts in SNOMED CT;
- Our approach differs from other quality assurance methods in that this approach can suggest remediations for the errors identified;
- We showed that identifying and analyzing small non-lattice graphlets in SNOMED CT with lexical patterns is a simple and effective quality assurance technique.

Acknowledgement

- The work was supported by University of Kentucky Center for Clinical and Translational Science (Clinical and Translational Science Award UL1TR001998).
- This work was also supported by the Intramural Research Program of the NIH, National Library of Medicine.

- The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.