The Subgraph Similarity Problem

Lorenzo De Nardo, Francesco Ranzato, Francesco Tapparo
Dipartimento di Matematica Pura ed Applicata, University of Padova, Italy

***Abstract*—Similarity is a well known weakening of bisimilarity where one system is required to simulate the other and vice versa. It has been shown that the subgraph bisimilarity problem, a variation of the subgraph isomorphism problem where isomorphism is weakened to bisimilarity, is NP-complete. We show that the subgraph similarity problem and some related variations thereof still remain NP-complete.**

***Index Terms*—Graph simulation, subgraph similarity problem, NP-completeness.**

I. Introduction

Bisimilarity, weak bisimilarity and similarity are well known behavioural equivalences that arise in every process calculus like CCS, π-calculus, ambient calculus, etc. [5]. These equivalences are also used as structural indexes to support efficient evaluation of query processing in graph-structured data, e.g. 1-index and A(k)-index for XML graphs [1], [4], [6], [9]. Dovier and Piazza [2] consider the so-called *subgraph bisimilarity problem*, a variant of the well-known NP-complete subgraph isomorphism problem where isomorphism is weakened to bisimilarity: this is the problem of identifying a subgraph G_2' of a graph G_2 bisimilar to a given graph G_1. The motivation for considering such a problem arises from data retrieval in query languages like G-log [8]. Dovier and Piazza prove that this problem remains NP-complete by means of a reduction of the Hamiltonian path problem (HP).

Similarity is a well known weakening of bisimilarity where one system is required to simulate the other and vice versa. We show that the *subgraph similarity problem* remains NP-complete and we still use a reduction of HP for proving this. On the other hand, weak bisimulation is a weakening of bisimulation that allows to bisimulate one step of a system by means of any finite number of steps. It turns out that weak bisimilarity is stronger than similarity. Thus, as a consequence, we also obtain that the *subgraph weak bisimilarity problem* is NP-complete.

II. Background

Let $R \subseteq X \times X$ and $S \subseteq X \times Y$ be binary relations. Then, $R^+ \subseteq X \times X$ denotes the transitive closure of R and $S^{-1} \subseteq Y \times X$ denotes the inverse relation $\{(y,x) \mid (x,y) \in S\}$. Given a graph G, $N(G)$ and $E(G)$ denote the sets of nodes and edges of G. The n-chain directed graph C_n, with $n \geq 1$, is $C_n = (\{x_1, ..., x_n\}, \rightarrow)$ where $x_i \rightarrow x_{i+1}$ for any $i \in [1, n-1]$.

The notions of (weak) simulation and (weak) bisimulation are given for labeled directed graphs in the context of process calculi or model checking, namely labeled transition systems when labels are attached to edges or Kripke structures when labels are attached to the nodes. The corresponding notions of (weak) simulation and (weak) bisimulation for unlabeled directed graphs are obtained as a particular case when one considers a single label. Let $G_1 = (N_1, \rightarrow_1)$ and $G_2 = (N_2, \rightarrow_2)$ be two directed graphs.

- A *simulation* of G_1 by G_2 is a relation $R \subseteq N_1 \times N_2$ such that: (1) R is total, i.e., for any $n \in N_1$ there exists $m \in N_2$ such that $(n,m) \in R$; (2) if $(n,m) \in R$ and $n \rightarrow_1 n'$ then there exists $m' \in N_2$ such that $(n', m') \in R$ and $m \rightarrow_2 m'$.
- A *weak simulation* of G_1 by G_2 is a relation $R \subseteq N_1 \times N_2$ such that: (1) R is total, i.e., for any $n \in N_1$ there exists $m \in N_2$ such that $(n,m) \in R$; (2) if $(n,m) \in R$ and $n \rightarrow_1 n'$ then there exists $m' \in N_2$ such that $(n', m') \in R$ and $m \rightarrow_2^+ m'$.
- G_1 and G_2 are (*weakly*) *bisimilar* when there exists a relation $R \subseteq N_1 \times N_2$, called (*weak*) *bisimulation*, such that: (1) R is a (weak) simulation of G_1 by G_2; (2) R^{-1} is a (weak) simulation of G_2 by G_1.
- G_1 and G_2 are (*weakly*) *similar* when there exist two relations $R \subseteq N_1 \times N_2$ and $S \subseteq N_2 \times N_1$ such that: (1) R is a (weak) simulation of G_1 by G_2; (2) S is a (weak) simulation of G_2 by G_1.

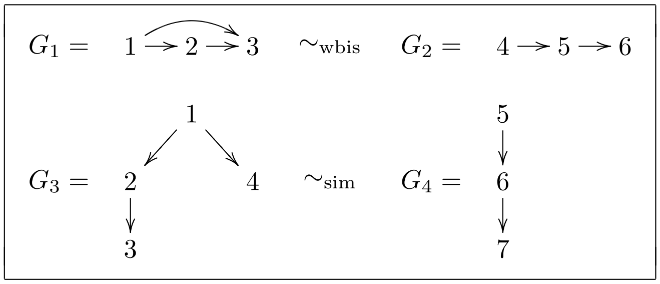
$G_1 \leq_{\mathrm{sim}} G_2$ denotes the fact that there exists a simulation of G_1 by G_2, so that G_1 and G_2 are similar when $G_1 \leq_{\mathrm{sim}} G_2$ and $G_2 \leq_{\mathrm{sim}} G_1$. Moreover, we denote, respectively, by $G_1 {\sim_{\mathrm{bis}}} G_2$, $G_1 {\sim_{\mathrm{wbis}}} G_2$, $G_1 {\sim_{\mathrm{sim}}} G_2$ and $G_1 {\sim_{\mathrm{wsim}}} G_2$ the fact that G_1 and G_2 are, respectively, bisimilar, weakly bisimilar, similar and weakly similar. Clearly, it turns out that $G_1 {\sim_{\mathrm{bis}}} G_2 \Rightarrow G_1 {\sim_{\mathrm{wbis}}} G_2$ and $G_1 {\sim_{\mathrm{sim}}} G_2 \Rightarrow G_1 {\sim_{\mathrm{wsim}}} G_2$.

Lemma II.1. *$G_1 {\sim_{\mathrm{wsim}}} G_2 \Rightarrow G_1 {\sim_{\mathrm{sim}}} G_2$ and $G_1 {\sim_{\mathrm{wbis}}} G_2 \Rightarrow G_1 {\sim_{\mathrm{sim}}} G_2$.*

Proof: Let $R \subseteq N_1 \times N_2$ be a weak simulation of G_1 by G_2. We consider the relation $R_+ \subseteq N_1 \times N_2$ obtained by adding to R the following pairs: for any $u, u' \in N_1$ and $v, v' \in N_2$ such that $u {\rightarrow_1} u'$, $(u,v) \in R$, $v {\rightarrow_2^+} v'$ and $(u', v') \in R$, if $v {\rightarrow_2^+} v'' {\rightarrow_2^+} v'$ then (u', v'') is added to R. Then, it is clear that R_+ is a simulation of G_1 by G_2. Moreover, if R is a weak bisimulation between G_1 and G_2, we consider the simulations R_+ and R_+^{-1} so that G_1 is similar to G_2. ■

Thus, $G_1 {\sim_{\mathrm{bis}}} G_2 \Rightarrow G_1 {\sim_{\mathrm{wbis}}} G_2 \Rightarrow G_1 {\sim_{\mathrm{sim}}} G_2 \Leftrightarrow G_1 {\sim_{\mathrm{wsim}}} G_2$. The following example shows that the first two implications are actually strict.

Example II.2. Let us consider the following graphs:



It turns out that $G_1 \sim_{\rm wbis} G_2$ by the weak bisimulation $R = \{(1,4),(2,5),(3,6)\}$ while there is no bisimulation between G_1 and G_2. On the other hand, we have that $G_3 \sim_{\rm sim} G_4$ by the simulations $R = \{(1,5),(2,6),(3,7),(4,6)\}$ and $S = \{(5,1),(6,2),(7,3)\}$ while there is no weak bisimulation between G_3 and G_4. □

The *subgraph bisimilarity* (respectively, *weak bisimilarity*, *similarity*) *problem*, denoted by $\mathrm{Bis}(G_1, G_2)$ (respectively, $\mathrm{WBis}(G_1, G_2)$, $\mathrm{Sim}(G_1, G_2)$), consists of deciding whether there exists a subgraph G_2' of G_2 such that $G_1 \sim_{\rm bis} G_2'$ (respectively, $G_1 \sim_{\rm wbis} G_2'$, $G_1 \sim_{\rm sim} G_2'$). The size of an instance of one of such problems is given by $|N_1| + |N_2| + |{\to_1}| + |{\to_2}|$.

III. The Subgraph Similarity Problem is NP-Complete

Dovier and Piazza [2] show that the subgraph bisimilarity problem Bis is NP-complete by reducing the directed Hamiltonian path problem HP to Bis. The proof is direct and basically depends on the fact that if a n-chain is bisimilar to a graph G with n nodes then G actually is isomorphic to the n-chain. We also reduce HP to Sim in order to prove that Sim is NP-hard: in this case the proof becomes less direct.

Let us first observe that Sim is in NP because $G_2' \sim_{\rm sim} G_1$ can be verified in polynomial time by using one polynomial-time simulation equivalence algorithm like that by Henzinger, Henzinger and Kopke [3] that runs in $O((|{\to_1}| + |{\to_2}|)(|N_1| + |N_2|))$. In fact, it is easy to show that similarity of two graphs G_1 and G_2 can be verified by a simulation equivalence algorithm on the disjoint union graph $G_1 \cup G_2 = (N_1 \cup N_2, {\to_1} \cup {\to_2})$.

Let us now show how HP can be reduced to Sim.

Lemma III.1. *If $G \sim_{\rm sim} C_n$ then C_n is isomorphic to a subgraph of G.*

Proof: Let us first show that G is an acyclic graph. Assume, by contradiction, that $a_1 \to a_2 \to \ldots a_k \to a_1$ is a cycle in G. Let R be the simulation relation of G by C_n. Then, there exists $x_j \in N(C_n)$ such that $(a_1, x_j) \in R$. Since $a_1 \to a_2$ and x_{j+1} is the unique successor of x_j, by simulation, we have that $(a_2, x_{j+1}) \in R$. Proceeding in this way, since $\{a_1, ..., a_k\}$ is a cycle, we would have that $(a_l, x_n) \in R$, for some $l \in [1, k]$. Since $a_l \to a_{l+1}$, we therefore would obtain that x_n must have a successor, which is a contradiction.

On the other hand, in a similar way, since $C_n \leq_{\rm sim} G$, it must be the case that G contains a path of length n. Since G is acyclic, the nodes in this path must be distinct, so that this path is indeed a n-chain. ■

Assume that G has n nodes. If G is similar to C_n then G contains a n-chain as subgraph but G is not necessarily isomorphic to C_n. By contrast, if G is bisimilar to C_n then G is isomorphic to C_n [2].

Theorem III.2. *Sim is NP-hard.*

Proof: Let us show that the Hamiltonian path problem HP can be reduced to Sim. Let G be a graph with $|N(G)| = n$. It turns out that the problem $\mathrm{HP}(G)$ is equivalent to $\mathrm{Sim}(C_n, G)$. On the one hand, if G admits an Hamiltonian path then such path is isomorphic to the n-chain C_n, so that G contains a subgraph which is similar to C_n. On the other hand, if G contains a subgraph G' which is similar to C_n then, by Lemma III.1, G' has a subgraph which is a n-chain and, since $|N(G)| = n$, this n-chain turns out to be an Hamiltonian path in G. To conclude we observe that this reduction can be done in polynomial time. ■

Corollary III.3. *WBis is NP-complete.*

Proof: Let G be a graph with $|N(G)| = n$. By the proof of Theorem III.2, $\mathrm{HP}(G) \Leftrightarrow \mathrm{Sim}(C_n, G)$. By the proof of [2, Theorem 1], $\mathrm{HP}(G) \Leftrightarrow \mathrm{Bis}(C_n, G)$. Moreover, $\mathrm{Bis}(C_n, G) \Rightarrow \mathrm{WBis}(C_n, G)$ trivially holds. On the other hand, by Lemma II.1, $\mathrm{WBis}(C_n, G) \Rightarrow \mathrm{Sim}(C_n, G)$. Thus, the Hamiltonian path problem HP can be reduced to WBis.

Let us finally observe that WBis is in NP because $G_2' \sim_{\rm wbis} G_1$ can be verified in polynomial time by first computing in polynomial time the transitive closure of G_2' and G_1 by Warshall's algorithm and then using a polynomial-time bisimulation algorithm like that by Paige and Tarjan [7] that runs in $O((|{\to_1}| + |{\to_2}|) \log(|N_1| + |N_2|))$. ■

Finally, it is worth remarking that Bis, WBis, Sim and WSim remain NP-complete problems also when considering the notions of (weak) bisimulation/simulation for labeled graphs, as the corresponding unlabeled problems can be reduced to them simply by considering a single label.

References

[1] P. Buneman, S. Davidson, M. Fernandez and D. Suciu. Adding structure to unstructured data. In *Proc. 6th ICDT*, LNCS 1186, pp. 336–350, 1997.

[2] A. Dovier and C. Piazza. The subgraph bisimulation problem. *IEEE Trans. Knowl. Data Eng.*, 15(4):1055-1056, 2003.

[3] M.R. Henzinger, T.A. Henzinger and P.W. Kopke. Computing simulations on finite and infinite graphs. In *Proc. 36th FOCS*, pp. 453–462, 1995.

[4] R. Kaushik, P. Shenoy, P. Bohannon and E. Gudes. Exploiting local similarity for indexing paths in graph-structured data. In *Proc. 18th ICDE*, pp. 129–140, 2002.

[5] R. Milner. *Communicating and Mobile Systems: the Pi-Calculus*. Cambridge University Press, 1999.

[6] T. Milo and D. Suciu. Index structures for path expressions. In *Proc. 7th ICDT*, LNCS 1540, pp. 277–295, 1999.

[7] R. Paige and R.E. Tarjan. Three partition refinement algorithms. *SIAM J. Comput.*, 16(6):973-989, 1987

[8] J. Paredaens, P. Peelman and L. Tanca. G-Log: A declarative graphical query language. *IEEE Trans. Knowl. Data Eng.*, 7(3):436-453, 1995.

[9] P. Ramanan. Covering indexes for XML queries: bisimulation – simulation = negation. In *Proc. 29th VLDB*, pp. 165–176, 2003.