Additional Material for the Paper GNN-Based Interactive Property Graph Repairs

Amedeo Pachera
Lyon1 University, CNRS Liris
Lyon, France
amedeo.pachera@univ-lyon1.fr

Angela Bonifati
Lyon1 University, CNRS Liris & IUF
Lyon, France
angela.bonifati@univ-lyon1.fr

Laks V.S. Lakshmanan
The University of British Columbia
Vancouver, Canada
laks@cs.ubc.ca

Andrea Mauri
Lyon1 University, CNRS Liris
Lyon, France
andrea.mauri@univ-lyon1.fr

PVLDB Reference Format:
Amedeo Pachera, Angela Bonifati, Laks V.S. Lakshmanan, and Andrea Mauri. Additional Material for the Paper GNN-Based Interactive Property Graph Repairs. PVLDB, 14(1): XXX-XXX, 2020.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at URL_TO_YOUR_ARTIFACTS.

1 TIME COMPLEXITY

Let $|V|$ be the number of nodes of the GRDG, $|F|$ the number of edges, $|U|$ the number of users, $C = |C_i|$ the size of candidates subgraphs per user and T the number of iterations of the algorithm.

Complexity of BuildCandidatesSet. Let Δ be the maximum degree of the GRDG. For user u_i, denote by CC_i the capacity and by D_v the node difficulties. Let $X_i := \min\left\{|V|, \left\lfloor \frac{CC_i}{\min_{v \in V} D_v} \right\rfloor\right\}$ be an upper bound on the number of nodes a candidate subgraphs for u_i can contain. Assume we start from S_i seeds of violations per user and run t_i iterations of 1-exchange local search (add/drop/swap one node) per seed, keeping at most $C_i \le S_i$ candidate subraphs. The greedy growth step performs at most X_i insertions, checking at most $O(X_i\Delta)$ neighboring nodes to select the one with maximal gain (cost $O(\log n)$), giving $T_{\text{greedy}}(i) = O\big(X_i\,\Delta\,\log n\big)$. One step of local search scans a 1-exchange neighborhood of size $O(X_i\Delta)$ with cached deltas; over t_i iterations this is $T(i) = O\big(t_i\,X_i\,\Delta\big)$. Hence the per-seed cost is $T_{\text{seed}}(i) = O\big(X_i\,\Delta\,(\log n + t_i)\big)$. With S_i seeds, the complexity for finding the candidate subgraphs for a user is $T_{\text{user}}(i) = S_i \cdot T_{\text{seed}}(i) = O\Big(S_i\,X_i\,\Delta\,(\log n + t_i)\Big)$,

Complexity of SubModularAssignment. For each user i, let C_i be the set of candidate subgraphs produced by BuildCandidatesSet. The total candidates for all the users are $M = \sum_{i=1}^{N} |C_i|$. and the total nodes accross the candidates are $Y = \sum_{i=1}^{N} \sum_{P \in C_i} |P|$. We can leverage an inverted index from each node v for the list of candidate items (i, P) that contain v (of size $O(M)$). Let $c_{\min} = \min_i c_i$ be the minimum cost among the users, and let X_{sel} be the number of selected candidates; clearly $X_{\text{sel}} \le \min\big\{M, |U|, \lfloor B/c_{\min} \rfloor\big\}$.

Computing initial gains $\sum_{v \in P} EQ(D_v, K_i)$ for all (i, P) and building the inverted index costs $O(S)$.

At each iteration of the greedy loop, we extract the current best feasible item in $O(\log M)$ time. Upon choosing $(i^\star, P^\star)$ we (i) remove all other patches of user $i^\star$ (at most $|C_{i^\star}|$ items overall across the run), and (ii) invalidate all candidates that overlap $P^\star$. Using the inverted index, step (ii) takes $O\big(\sum_{v \in P^\star} \Gamma_v\big)$ time, where Γ_v is the number of candidate items containing node v. Because subgraphs are enforced disjoint, each node is invalidated at most once; hence across the whole run $\sum_{\text{selected } P^\star} \sum_{v \in P^\star} \Gamma_v = \sum_v \Gamma_v = Y$. The total number of operations on the lists of candidates is $O(M + X_{\text{sel}})$, so heap work is $O\big((M + X_{\text{sel}}) \log M\big)$.

Summing up, $T_{\text{assignment}} = O\big(Y + (M + X_{\text{sel}}) \log M\big)$. Under uniform parameters ($|C_i| = L$ candidates per user, each of size at most X), we have $M = |U|L$ and $X \le |U|LX$, giving

$$T_{\text{assignment}} = O\big(|U|LX + (|U|L + X_{\text{sel}}) \log(|U|L)\big).$$

Complexity of the Final Outer Loop. Let T be the maximum number of Lagrangian iterations. The remaining outer-loop work per iteration is lightweight: (i) computing the current primal value Z is $O\big(\sum_{i \in U'} |P_i|\big) \le O(|V|)$, (ii) the dual quantity UB_t is $O(1)$ once Z and $\sum_{i \in U'} c_i$ are known, (iii) the subgradient update and projection of λ are $O(1)$, and (iv) updating the incumbent $(U^\star, \{P_i^\star\})$ is $O(1)$.

Hence the worst-case total time (without early stopping) is $T_{\text{total}} = T \cdot \Big(|U| \cdot T_{\text{user}} + T_{\text{assignmnet}} + O(|V|)\Big)$. Therefore, T_{total} is

$$O\Big(T \cdot \big(|U|\,S\,X\,\Delta\,(\log|V| + t) + |U|LX + (|U|L + X_{\text{sel}}) \log(|U|L) + |V|\big)\Big).$$

2 CONSTRAINTS

In this section, we list the constraint used for each dataset, in a datalog-like format.

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

2.1 FAERS (Real)

A drug cannot be primary suspect and secondary suspect of the same case

$$\begin{aligned}\phi_1 :=&(Case(X) \wedge Drug(Y) \wedge Caze(Z) \\ &\wedge (IS\_PRIMARY\_SUSPECT(X,Y) \\ &\wedge IS\_SECONDARY\_SUSPECT(Y,Z), \\ &id(X) = id(Y) \rightarrow \bot)\end{aligned}$$

A drug cannot be primary suspect of itself

$$\begin{aligned}\phi_1 :=&(Drug(X) \wedge Drug(Y) \\ &\wedge (IS\_PRIMARY\_SUSPECT(X,Y) \\ &id(X) = id(Y) \rightarrow \bot)\end{aligned}$$

A case cannot fall under two different age groups

$$\begin{aligned}\phi_3 :=&(Case(X) \wedge AGE\_GROUP(Y) \wedge AGE\_GROUP(z) \\ &\wedge (FALLS\_UNDER(X,Y) \wedge FALLS\_UNDER(X,Z), \\ &id(Z) \neq id(Y) \rightarrow \bot)\end{aligned}$$

A drug cannot be prescribed to a child

$$\begin{aligned}\phi_3 :=&(Terapy(X) \wedge Drug(Y) \wedge Case(Z) \wedge AGE\_GROUP(W) \\ &\wedge (PRESCRIBED(X,Y) \wedge RECEIVED(Y,Z) \\ &\wedge FALLS\_UNDER(Z,W) \\ &W.ageGroup = \text{"}Child\text{"} \rightarrow \bot)\end{aligned}$$

2.2 LDBC Finbench

An account cannot transfer money if an account is blocked

$$\begin{aligned}\phi_1 :=&(Account(X) \wedge Account(Y) \\ &\wedge (Transfer(X,Y), X.isBLocked = True \rightarrow \bot)\end{aligned}$$

A guarantor needs to have more than 5M in the bank account

$$\begin{aligned}\phi_2 :=&(Person(X) \wedge Person(Y) \wedge Account(Z) \\ &\wedge (Guarantee(X,Y) \wedge Own(Y,Z), \\ &Z.balance < 500000000 \wedge id(X) \neq id(Y) \rightarrow \bot)\end{aligned}$$

A person cannot be guarantor of themselves

$$\begin{aligned}\phi_3 :=&(Account(X) \wedge Account(Y) \\ &\wedge (Guarantee(X,Y), id(X) = id(Y) \rightarrow \bot)\end{aligned}$$

Interest Rates for company need to be greater than 0.2

$$\begin{aligned}\phi_4 :=&(Company(X) \wedge Loan(Y) \\ &\wedge (Apply(X,Y), Y.interestRate < 0.2 \rightarrow \bot)\end{aligned}$$

2.3 ICIJ (Real)

An officer of a company cannot register the same address that an Entity has registered

$$\begin{aligned}\phi_1 :=&(Officer(X) \wedge Entity(Y) \wedge Address(Y) \wedge Officer(W) \\ &\wedge officer\_of(X,Y) \wedge registered\_address(Y,Z) \\ &\wedge registered\_address(W,Z), \\ &id(X) = id(W) \rightarrow \bot)\end{aligned}$$

The registered address of an entity must be the same as the one of the address

$$\begin{aligned}\phi_2 :=&Address(X) \wedge Company(Y) \\ &\wedge (registered\_address(X,Y), \\ &X.country\_codes \neq Y.country\_codes \rightarrow \bot)\end{aligned}$$

An intermediary cannot be a shareholder in the same company for which they act as an intermediary

$$\begin{aligned}\phi_1 :=&(Entity(X) \wedge Intermediary(Y) \wedge Entity(Z) \\ &\wedge intermediary\_of(Y,X) \wedge shareholder\_of(Y,Z), \\ &id(X) = id(Z) \rightarrow \bot)\end{aligned}$$

A shareholder of a company needs to have at least 1700000 dollars

$$\begin{aligned}\phi_2 :=&(Officer(X) \wedge Company(Y) \\ &\wedge shareholder\_of(X,Y), X.networth < 1700000 \rightarrow \bot)\end{aligned}$$

2.4 LDBC SNB

A comment can't be created before its post.

$$\begin{aligned}\phi_1 :=& (Comment(X) \wedge Post(Y) \wedge TO(X,Y) \\ &\wedge\ X.creationTime < Y.creationTime \rightarrow \bot)\end{aligned}$$

Under-age people can't be members of a forum with a higher requirement.

$$\begin{aligned}\phi_2 :=& (Person(X) \wedge Forum(Y) \wedge HAS\_MEMBER(Y,X) \\ &\wedge\ X.age < Y.ageRequirement \rightarrow \bot)\end{aligned}$$

A person must work at an organisation located in the same place they live.

$$\begin{aligned}\phi_3 :=& (Place(X) \wedge Place(Y) \wedge Person(Z) \wedge Organisation(W) \\ &\wedge\ LIVES\_IN(Z,X) \wedge WORK\_AT(X,W) \\ &\wedge\ LOCATED\_IN(W,Y) \wedge id(X)! = id(Y) \rightarrow \bot)\end{aligned}$$

A person can't like their own post.

$$\begin{aligned}\phi_4 :=& (Post(X) \wedge Person(Y) \wedge CREATED(Y,X) \\ &\wedge\ LIKES(X,Y) \rightarrow \bot)\end{aligned}$$

REFERENCES