Approximate Querying on Property Graphs
Companion Appendix

Stefania Dumbrava[1], Angela Bonifati[2], Amaia Nazabal Ruiz Diaz[2], and Romain Vuillemot[3]

[1] ENSIIE Évry & CNRS Samovar
stefania.dumbrava@ensiie.fr
[2] University of Lyon 1 & CNRS LIRIS
{angela.bonifati,amaia.nazabal-ruiz-diaz}@univ-lyon1.fr
[3] École Centrale Lyon & CNRS LIRIS
romain.vuillemot@ec-lyon.fr

We complement the information in the paper, as follows. In Section 1, we give the formulas for the precomputed properties used to estimate counting RPQs. In Section 2, we establish intermediate results characterizing the computed graph summarization. Finally, in Section 3, we detail the NP-completeness proof regarding summarization optimality, under our algorithm's conditions.

1 Precomputed Properties

Formulas for relevant precomputed properties are given in Fig.1. Each *supernode*, v^*, comprises all subgrouping vertices and edges, $\mathcal{G}_i^* = (V_i^*, E_i^*)$, formed during the *grouping phase*. **Note**: l_c indicate cross-edge labels and l_i, inner-edges ones.

Property	**Definition**
$VWeight(v^*)$	$\lvert V_i^* \rvert$
$EWeight(v^*)$	$\lvert E_i^* \rvert$
$LPercent(v^*, l)$	$\lvert \{e \in E_i^* \mid e = l(\_,\_)\} \rvert / EWeight(v^*)$
$LReach(v^*, l)$	$\lvert \{(v_1, v_2) \in V_i^* \times V_i^* \mid l^+(v_1, v_2) \in \mathcal{G}_i^*\} \rvert$
$V_F(v^*, l, d)$	$\lvert \{v \mid v \in v^* \wedge \exists e, e(\_,\_) \in E \setminus E_i^* \wedge e.d = v\} \rvert$
$LPart(v^*, l_c, l_i, d_c, d_i)$	$TReach(v^*, l_i, d_i) / \lvert V_F(v^*, l_c, d_c) \rvert$
$EWeight(e^*)$	$\lvert \{e \in E \mid e \in e^*\} \rvert$
$LPercent(\hat{v}, l)$	$((\sum_{v^* \in \hat{v}} LPercent(v^*, l) * EWeight(v^*)) / \sum_{v^* \in \hat{v}} EWeight(v^*)$

Fig. 1: Precomputed Graph Summary Properties

2 Grouping Characterization

We henceforth denote $\Phi = GROUPING(\mathcal{G})$ and name each $\mathcal{G}' \in \Phi$, a $\mathcal{G}$-*grouping* and each $\mathcal{G}'' \in \mathcal{G}'$, a $\mathcal{G}'$-*subgrouping*. Note that Φ is not unique, as, for $l_1, l_2 \in \Lambda(\mathcal{G})$, s.t $\#l_1 = \#l_2$, we arbitrarily order l_1 and l_2 in $\overrightarrow{\Lambda(\mathcal{G})}$.

Definition 1 (Non-Trivial (Sub)Groupings). *A $\mathcal{G}$-grouping, $\mathcal{G}' = (V', E')$, is called* trivial, *if $\mathcal{G}' = \mathcal{G}$ or $E' = \emptyset$, and* non-trivial, *otherwise. A $\mathcal{G}'$-subgrouping, $\mathcal{G}'' = (V'', E'')$, is called* trivial, *if $E'' = \emptyset$, and* non-trivial, *otherwise.*

Lemma 1 (Non-Trivial Grouping Properties). *Let $\mathcal{G}'$ be a non-trivial $\mathcal{G}$-grouping. The following hold.* ***P1:*** *For any non-trivial $\mathcal{G}'$-subgrouping, $\mathcal{G}''$, there exists $l'' \in \Lambda(\mathcal{G}')$, s.t $\lambda(\mathcal{G}') = l''$.* ***P2:*** *For any non-trivial distinct $\mathcal{G}'$-subgroupings, $\mathcal{G}''_1$, $\mathcal{G}''_2$: a) $\lambda(\mathcal{G}''_1) = \lambda(\mathcal{G}''_2)$ and b) $\mathcal{G}''_1$ and $\mathcal{G}''_2$ are edge-wise disjoint.*

Proof. **P1** is provable by contradiction. If $\nexists l''$, $l'' \in \Lambda(\mathcal{G}')$, s.t $\lambda(\mathcal{G}') = l''$, then $E' = \emptyset$, contradicting the non-triviality of $\mathcal{G}'$. **P2.a)** holds by construction and **P2.b)**, by contradiction. Assume $\mathcal{G}''_1 \cap \mathcal{G}''_2 \neq \emptyset$; then, $\mathcal{G}''_1$ and $\mathcal{G}''_2$ share at least a node, which is impossible by construction. □

We characterize the *GROUPING* algorithm, based on the following remarks.

Lemma 2 (Subgrouping Maximal Label-Connectivity). *For each $\mathcal{G}_i \in \Phi$, each of its* maximally weakly connected *components, $\mathcal{G}^*_i \in \mathcal{G}_i$, is also* maximally label-connected *on l, where $\#l = \max\limits_{l \in \Lambda(\mathcal{G}_i)} (\#l)$.*

Proof. By construction, we know that, if $\mathcal{G}^*_i \in \mathcal{G}_i$, there exists $l' \in \Lambda(\mathcal{G})$, such that $\lambda(\mathcal{G}^*_i) = l'$. Assume that $l' \neq l$. By definition, there exists at least one l-labeled edge in E^*_i. Since $\mathcal{G}^*_i$ is maximally label-connected on l', then each such edge connects vertices also connected by an edge labeled l'. As $\#l \geq \#l'$, then there exists at least one pair of vertices in V^*_i connected by more edges labeled l than l'. Hence, $\lambda(\mathcal{G}^*_i) = l$, contradicting the hypothesis. □

Theorem 1 (*GROUPING* Properties). *If $|V| \geq 1$, then:*
P1: $\forall \mathcal{G}_i \in \Phi, V_i \neq \emptyset$
P2: $\forall \mathcal{G}_i, \mathcal{G}_j \in \Phi$, *where* $i \neq j$, $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset$
P3: $\bigcup\limits_{i \in [1,k]} V_i = V$ *and* $\bigcup\limits_{i \in [1,k]} E_i \subseteq E$
P4: $\Phi = \{\mathcal{G}_i = (V_i, E_i) \subseteq \mathcal{G} \mid i \in [1, |\Lambda(\mathcal{G})| + 1]\}$

Proof. **P1**, **P2**, **P3** trivially hold. We prove **P4**. If $E = \emptyset$, $\Phi = \{\mathcal{G}\}$. Otherwise, there exists $l \in \overrightarrow{\Lambda(\mathcal{G})}$ and $\mathcal{G}_i \in \Phi$, such that $\lambda(\mathcal{G}_i) = l$. Assume $\Phi > |\Lambda(\mathcal{G})| + 1$. At least two groupings, $\mathcal{G}_i, \mathcal{G}_j$, with the same most frequently occurring label, l, exist. As $|\mathcal{G}_i| \geq 1$, $|\mathcal{G}_j| \geq 1$, each contains a maximally weakly connected component, $\mathcal{G}'_i$, $\mathcal{G}'_j$. From Lemma 2, $\lambda(\mathcal{G}'_i) = \lambda(\mathcal{G}'_j)$, contradicting $\mathcal{G}_i \cap \mathcal{G}_j \neq \emptyset$. □

3 Optimal Summary Intractability

Theorem 2. *Let* MinSummary *be the problem that, for a graph $\mathcal{G}$ and an integer $k' \geq 2$, decides if there exists a label-driven partitioning Φ of $\mathcal{G}$, $|\Phi| \leq k'$, s.t χ_Λ is a* valid summarization. *MinSummary is NP-complete, even for undirected graphs, $|\Lambda(\mathcal{G})| \leq 2$ and $k' = 2$.*

Proof. We establish the result in two steps. First, **MinSummary is in NP**. We construct a valid *summarization function*, χ_Λ, as a witness. For a graph partitioning in k *subgraphs*, one can verify in polynomial time if two vertices are reachable by a given labeled-constrained path and decide if their assignation to the same or to different HNs is valid. Second, **MinSummary is NP-hard**. We reduce the **MinSummary** problem to **IndSet**, i.e., the NP-complete problem of establishing whether an undirected graph contains K independent vertices, for an arbitrary K. We prove **IndSet** $\leq_p$ **MinSummary**. Let $\mathcal{G} = (V, E)$ be an **IndSet** instance, where $\mathcal{G}$ is undirected, $|V| = n \geq 2$, $|E| = m$, $\Lambda(\mathcal{G}) = \{l_1\}$. We consider a polynomial reduction function, f, s.t $f(\mathcal{G}) = \mathcal{G}'$, $\mathcal{G}' = (V', E')$ (see Fig. 2), $\{v'_1, v'_2, v'_3\} \subset V'$, $\Lambda(\mathcal{G}) = \{l_1, l_2\}$, and $\tilde{\mathcal{G}} \subset \mathcal{G}$, where $\tilde{\mathcal{G}}$ is obtained from $\mathcal{G}$, by adding, between each pair of vertices connected with an l_1-labeled edge, n more l_1-labeled edges. Let $\mathcal{G}'$ contain three paths of length k, between v'_1 and v'_2 (one, l_1-labeled, and two, l_2-labeled) and two paths of length n, between v'_2 and v'_3, of each color. Let $K \geq 0$ be the number of independent vertices in $\mathcal{G}$. In $\mathcal{G}'$, $\#l_1 \geq (n+1)(n-K-1) + 2k + n$ and $\#l_2 = 2n + k$. $l_2 = \max_{l \in \mathcal{G}'}(\#l) \Rightarrow K \geq \frac{n^2-n-1+k}{n+1} \geq 1$. We show: $\mathcal{G}$ satisfies **IndSet** $\Leftrightarrow$ $\mathcal{G}'$ satisfies **MinSummary**.

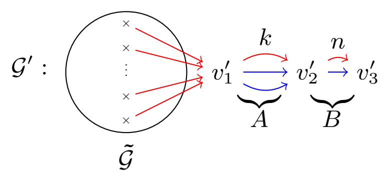

Fig. 2: $\mathcal{G}'$ Construction

$\Rightarrow$ Let $\mathcal{G}$ satisfy **IndSet**. We can thus choose a set of independent vertices $S \subset V$, $|S| = k$. Let $\mathcal{G}_2$ be the $\mathcal{G}'$-subgraph induced by $S \cup A \cup B$. It is *maximally l_2-connected* and contains $2k + n$ edges labeled l_2 and $2k + n$ edges labeled l_1, i.e., $\lambda(\mathcal{G}_2) = l_2$. Let $\mathcal{G}_1$ be the $\mathcal{G}'$-subgraph induced by $V \setminus S$. It is *maximally l_1-connected* and contains $(n + 1)m$ edges, all labeled l_1; hence, $\lambda(\mathcal{G}_1) = l_1$. $\Phi = \{\mathcal{G}_1, \mathcal{G}_2\}$ is a valid summarization of $\mathcal{G}'$, as $l_1 = \max_{l \in \mathcal{G}_1}(\#l)$ and $l_2 = \max_{l \in \mathcal{G}_2}(\#l)$. $\mathcal{G}'$ satisfies **MinSummary**.
$\Leftarrow$ Let $\mathcal{G}'$ satisfy **MinSummary**. We can thus compute a $\mathcal{G}$-partitioning, Φ, that is a *valid summarization*, where $|\Phi| \leq 2$. If $\Phi = 2$, then there exist two distinct $\mathcal{G}'$-subgraphs, $\mathcal{G}_1$, $\mathcal{G}_2$, where $\Phi = \{\mathcal{G}_1, \mathcal{G}_2\}$. As $\#l_1 = (n+1)m+2k+n \geq 2n+k = \#l_2$ in $\mathcal{G}'$, one of the subgraphs $\mathcal{G}_1$, $\mathcal{G}_2$, should be s.t all of its components are *maximally l_1-connected*. Let that subgraph be $\mathcal{G}_1$. Hence, $\mathcal{G}_1 \cap \tilde{\mathcal{G}}$ contains all vertices connected by a l_1-labeled edge. We denote by $\tilde{V_1}$ the set of vertices in $\mathcal{G}_1 \cap \tilde{\mathcal{G}}$. The set of vertices in $\mathcal{G}_1$ is thus $\tilde{V_1} \cup A \cup B$. As Φ has to be a valid summarization, the set of vertices in $\mathcal{G}_2$ is V_2, where $V_2 = V' \setminus (\tilde{V_1} \cup A \cup B)$. We can thus choose the set of independent vertices of size K in $\mathcal{G}$ to be $S = V_2$. If $|\Phi| = 1$, $\Phi = \{\mathcal{G}'\}$ must be a *valid summarization* of $\mathcal{G}'$. As $\mathcal{G}'$ is *maximally l_2-connected*, it must hold that $l_2 = \max_{l \in \mathcal{G}'}(\#l)$. Hence, $K \geq 1$ and we can choose the set of independent vertices in $\mathcal{G}$ to be $S = V' \cap V$. Thus, $\mathcal{G}$ satisfies **IndSet**. $\square$