Full Proof of Theorem 1

March 22, 2020

1 The Proof

The key idea of our proof is to show the existence of a path between some query entity in an optimum solution and a certificate for its representability such that: for every entity in the path, its priority is not smaller than the score of that optimum solution. Consequently, as search starts from each query entity, it is impossible that the algorithm produces a sub-optimum solution without visiting the above certificate via the path.

A Formal Proof:

We prove by contradiction.

Assume Algorithm 1 returns a sub-optimum solution $Q_{\max}$ whose score is smaller than that of an optimum solution denoted by Q_{opt}. Let $c \in \mathbb{V}$ be a certificate for the representability of Q_{opt}. Before Algorithm 1 is terminated, c has never been checked by QMaxCertWith because otherwise, Q_{opt} instead of $Q_{\max}$ would be returned. However, we can prove the existence of a particular path between c and some query entity $q \in Q_{\text{opt}}$, called a key path, such that: for every entity e in this path, $pr_{e|q} \geq \texttt{score}(Q_{\text{opt}}) > \texttt{score}(Q_{\max})$ holds. With this key path, the algorithm is impossible to return $Q_{\max}$ without visiting c via this path and finding Q_{opt} by QMaxCertWith, leading to a contradiction.

To show the existence of such a key path, consider an ERG G' constructed by merging shortest paths in the following way to connect Q_{opt}, where c is a certificate for Q_{opt} defined in Lemma 1. (1) When D is even, for each query entity $q \in Q_{\text{opt}}$, we choose a shortest path between q and c in G. All of these paths are merged into a connected subgraph G', which clearly satisfies $\texttt{diam}(G') \leq D$ according to Lemma 1. In particular, if there are more than one shortest path between two vertices, we consistently choose one of them in a deterministic way to avoid cycles. It can be determined with the help of a fixed order of the arcs in $\mathbb{E}$, e.g., in alphabetical order of their identifiers. This ensures that G' is a tree and hence is minimal. Therefore, G' is an ERG. (2) When D is odd, we construct a minimal connected subgraph G' in a similar way. In particular, when $q \in Q_{\text{opt}}$ is $\lceil \frac{D}{2} \rceil$ hops away from c, we firstly choose a shortest path between q and c' in G (where c' is defined in Lemma 1), and then merge that path with the arc between c' and c to form a shortest path between q and c. This ensures that $\texttt{diam}(G') \leq D$. Therefore, G' is an ERG.

Now we prove that: (1) at least one of the above shortest paths is not longer than $\lfloor \frac{D}{2} \rfloor$, and (2) this shortest path is a key path.

(1) When D is even, all of those shortest paths are not longer than $\lceil \frac{D}{2} \rceil = \lfloor \frac{D}{2} \rfloor$. When D is odd, assume on the contrary that all of them are longer than $\lfloor \frac{D}{2} \rfloor$. According to Lemma 1, all the query entities need to be considered for Condition 2 of the lemma, and hence all of those shortest paths pass through c' which is a neighbor of c. That contradicts the minimality of G' in the definition of ERG because vertex c and the arc between c and c' can be removed from G' to obtain a proper subgraph of G' that is still a well-defined ERG.

(2) Let p be a path not longer than $\lfloor \frac{D}{2} \rfloor$ proved in (1), which connects $q \in Q_{\text{opt}}$ and c. For every entity e in p and every query entity $q' \in (Q_{\text{opt}} \setminus \{q\})$, $\texttt{dist}(e, q) + \texttt{dist}(e, q')$ is not larger than the sum of the length of p and the length of a shortest path between c and q', which in turn is not larger than $\lfloor \frac{D}{2} \rfloor + \lceil \frac{D}{2} \rceil = D$. So $q' \in Q_{\text{ub}}(e|q)$ and $Q_{\text{opt}} \subseteq Q_{\text{ub}}(e|q)$, and hence $pr_{e|q} = \texttt{score}(Q_{\text{ub}}(e|q)) \geq \texttt{score}(Q_{\text{opt}}) > \texttt{score}(Q_{\text{max}})$, so p is a key path.

■