Querying Shortest Path Distance with Bounded Error in Large Graphs

Miao Qiao, Hong Cheng, Jeffrey Xu Yu

Department of Systems Engineering & Engineering Management
The Chinese University of Hong Kong

July 19, 2011

Roadmap

- Related work
- Problem Statement
- Basic algorithms
- Graph Partitioning-based Heuristic
- Experiments

Roadmap

- Related work
- Problem Statement
- Basic algorithms
- Graph Partitioning-based Heuristic
- Experiments

Related Work

Graph embedding technique have been widely used in calculating shortest path in

- Road networks [10, 5]

Related Work

Graph embedding technique have been widely used in calculating shortest path in

- Road networks [10, 5]
- Social networks and web graphs [8, 7, 9, 2]

Related Work

Graph embedding technique have been widely used in calculating shortest path in

- Road networks [10, 5]
- Social networks and web graphs [8, 7, 9, 2]
- Internet [1, 6].

Related Work

Graph embedding technique have been widely used in calculating shortest path in

- Road networks [10, 5]
- Social networks and web graphs [8, 7, 9, 2]
- Internet [1, 6].

Related Work

Graph embedding technique have been widely used in calculating shortest path in

- Road networks [10, 5]
- Social networks and web graphs [8, 7, 9, 2]
- Internet [1, 6].

The major difference between them lie in three aspects

- Reference Node selection [7]
  Potamias et al. betweenness centrality.

Related Work

Graph embedding technique have been widely used in calculating shortest path in

- Road networks [10, 5]
- Social networks and web graphs [8, 7, 9, 2]
- Internet [1, 6].

The major difference between them lie in three aspects

- Reference Node selection [7]
  Potamias et al. betweenness centrality.
- Reference Node organization [1, 5]
  Kriegel et al. hierarchical embedding.

Related Work

Graph embedding technique have been widely used in calculating shortest path in

- Road networks [10, 5]
- Social networks and web graphs [8, 7, 9, 2]
- Internet [1, 6].

The major difference between them lie in three aspects

- Reference Node selection [7]
  Potamias et al. betweenness centrality.
- Reference Node organization [1, 5]
  Kriegel et al. hierarchical embedding.
- An error bound or not. [11, 9, 2]
  Thorup and Zwick et al. $(2k-1)$-approximation with $O(kn^{1+1/k})$ memory.

- Related work
- Problem Statement
- Basic algorithms
- Graph Partitioning-based Heuristic
- Experiments

Problem Statement

Problem (Distance Estimation with a Bounded Error)

Input: a graph G and a user-specified error bound ϵ, for query (s, t)
Output: a estimated shortest distance $\widehat{D}(s, t)$, with error

$$|\widehat{D}(s, t) - D(s, t)| \leq \epsilon$$

.

Problem Statement

Problem (Distance Estimation with a Bounded Error)

Input: a graph G and a user-specified error bound ϵ, for query (s, t)
Output: a estimated shortest distance $\widehat{D}(s, t)$, with error

$$|\widehat{D}(s, t) - D(s, t)| \leq \epsilon$$

.

The question to discuss :

- How to select the minimum number of reference nodes to ensure the error bound ϵ?

- Problem Statement
- Basic algorithms
- Graph Partitioning-based Heuristic
- Experiments

- Problem Statement
- Basic algorithms
- Graph Partitioning-based Heuristic
- Experiments

- Problem Statement
- Basic algorithms
  - Reference Node Selection
  - Shortest Path Distance Estimation
  - Error Bound Analysis
- Graph Partitioning-based Heuristic
- Experiments

- Problem Statement
- Basic algorithms
  - Reference Node Selection
  - Shortest Path Distance Estimation
  - Error Bound Analysis
- Graph Partitioning-based Heuristic
- Experiments

Reference Node Selection

Definition (Coverage)

Given a graph $G = (V, E, w)$ and a radius c, a vertex $v \in V$ is covered by a reference node r if $D(r, v) \leq c$.

Reference Node Selection

Definition (Coverage)

Given a graph $G = (V, E, w)$ and a radius c, a vertex $v \in V$ is covered by a reference node r if $D(r, v) \leq c$.

Problem (Coverage-based Reference Node Selection)

Input: a graph $G = (V, E, w)$ and a radius c
Output: R^*, $R^* = \arg\min_{R \subseteq V} |R|$, s.t. $\forall v \in V - R^*$, v is covered by at least one reference node from R^*.

Reference Node Selection

Definition (Gain Function)

The gain function over a set of reference nodes R is defined as

$$g(R) = |\cup_{r \in R} C_r| - |R|$$

Reference Node Selection

Definition (Gain Function)

The gain function over a set of reference nodes R is defined as

$$g(R) = |\cup_{r\in R} C_r| - |R|$$

$$\max_R g(R) = \max_R(|\bigcup_{r\in R} C_r| - |R|) = |V| - \min_R |R| = g(R^*)$$

Reference Node Selection

Definition (Gain Function)

The gain function over a set of reference nodes R is defined as

$$g(R) = |\cup_{r \in R} C_r| - |R|$$

$$\max_R g(R) = \max_R(|\bigcup_{r \in R} C_r| - |R|) = |V| - \min_R |R| = g(R^*)$$

g is a submodular function and in general maximizing a submodular function is NP-hard[4].

Reference Node Selection

Definition (Gain Function)

The gain function over a set of reference nodes R is defined as

$$g(R) = |\cup_{r \in R} C_r| - |R|$$

$$\max_R g(R) = \max_R(|\bigcup_{r \in R} C_r| - |R|) = |V| - \min_R |R| = g(R^*)$$

g is a submodular function and in general maximizing a submodular function is NP-hard[4].
A greedy algorithm that select $r_k \in R$ in the k-th iteration:

$$r_k = \arg \max_{r \in V \setminus R_{k-1}} g(R_{k-1} \cup \{r\}) - g(R_{k-1}) \quad (1)$$

Reference Node Selection

Sparse part of graph or isolated vertices may cause $|R|$ unnecessarily large.

Reference Node Selection

Sparse part of graph or isolated vertices may cause $|R|$ unnecessarily large.

Definition (Cover Ratio)

The percentage of vertices in V are covered by R.

Shortest Path Distance Estimation

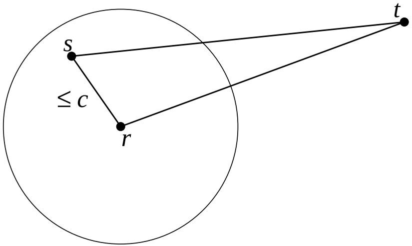

Figure: Distance Estimation

$$D(s,t) \leq \widehat{D}_U(s,t) = \min_{r \in \mathcal{R}} (D(s,r) + D(r,t)) \quad (2)$$

Error Bound Analysis

Theorem

Given any query (s, t), *error bound* ϵ, *with the coverage radius* $c = \frac{\epsilon}{2}$ *and* $err(s, t) = |\widehat{D}(s, t) - D(s, t)|$,

$$P(err(s, t) \leq \epsilon) \geq 1 - (1 - CR)^2$$

.

Error Bound Analysis

Theorem

Given any query (s, t)*, error bound* ϵ*, with the coverage radius* $c = \frac{\epsilon}{2}$ *and* $err(s,t) = |\widehat{D}(s,t) - D(s,t)|$,

$$P(err(s,t) \leq \epsilon) \geq 1 - (1 - CR)^2$$

.

When $CR = 0.8$, the bound is satisfied with a probability $P(err(s,t) \leq \epsilon) \geq 0.96$.

- Problem Statement
- Basic algorithms
- Graph Partitioning-based Heuristic
    - Partitioning-based Reference Node Embedding
    - Partitioning-based Shortest Distance Estimation
    - Error bound Analysis
- Experiments

Partitioning-based Reference Node Embedding

- Select reference node R as previous section described.

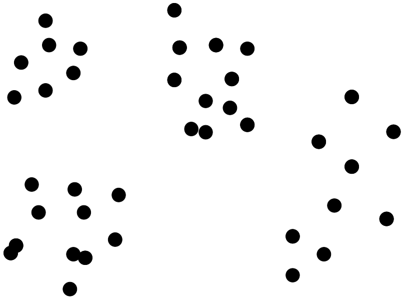

Figure: Distance Estimation in RN-partition

Partitioning-based Reference Node Embedding

- Select reference node R as previous section described.

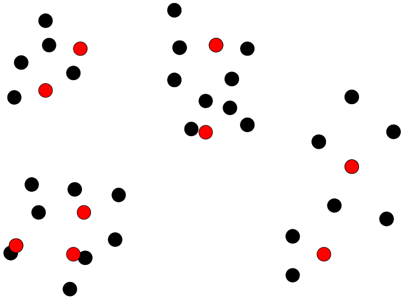

Figure: Distance Estimation in RN-partition

Partitioning-based Reference Node Embedding

- Select reference node R as previous section described.
- Use KMETIS[3] to partition the graph into K clusters $C_1, \ldots, C_K$.

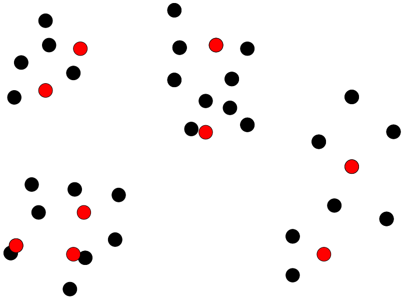

Figure: Distance Estimation in RN-partition

Partitioning-based Reference Node Embedding

- Select reference node R as previous section described.
- Use KMETIS[3] to partition the graph into K clusters $C_1, \ldots, C_K$.

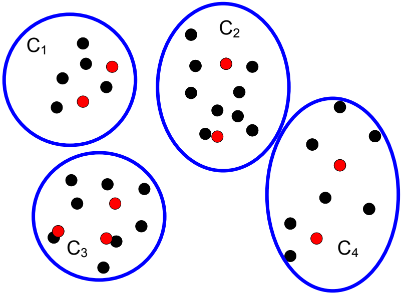

Figure: Distance Estimation in RN-partition

Partitioning-based Reference Node Embedding

- Select reference node R as previous section described.
- Use KMETIS[3] to partition the graph into K clusters $C_1, \ldots, C_K$.
- Assign R into K set R_i with

$$R_i = \{r | r \in R \text{ and } r \in C_i\}.$$

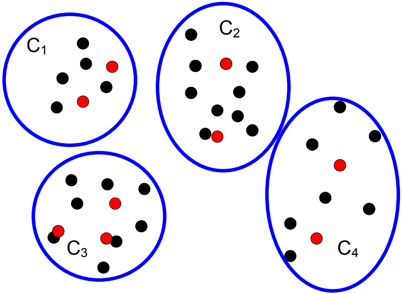

Figure: Distance Estimation in RN-partition

Partitioning-based Reference Node Embedding

- Select reference node R as previous section described.
- Use KMETIS[3] to partition the graph into K clusters $C_1, \ldots, C_K$.
- Assign R into K set R_i with

$$R_i = \{r | r \in R \text{ and } r \in C_i\}.$$

- Compress R_i as a super node SN_i.

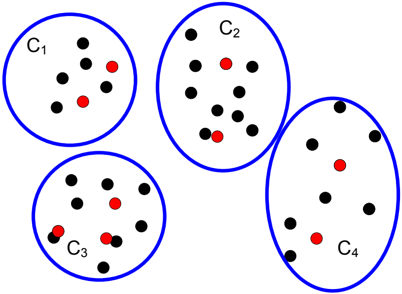

Figure: Distance Estimation in RN-partition

Partitioning-based Reference Node Embedding

- Select reference node R as previous section described.
- Use KMETIS[3] to partition the graph into K clusters $C_1, \ldots, C_K$.
- Assign R into K set R_i with

$$R_i = \{r | r \in R \text{ and } r \in C_i\}.$$

- Compress R_i as a super node SN_i.
- Compute the single source shortest paths from SN_i to every vertex $v \in V$ where $i \in [1..K]$.

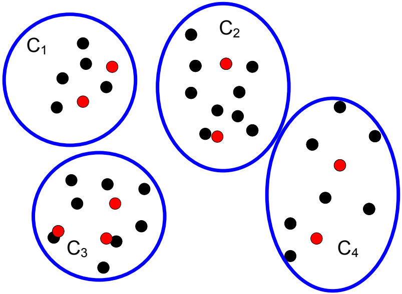

Figure: Distance Estimation in RN-partition

Partitioning-based Reference Node Embedding

- Select reference node R as previous section described.
- Use KMETIS[3] to partition the graph into K clusters $C_1, \ldots, C_K$.
- Assign R into K set R_i with

$$R_i = \{r | r \in R \text{ and } r \in C_i\}.$$

- Compress R_i as a super node SN_i.
- Compute the single source shortest paths from SN_i to every vertex $v \in V$ where $i \in [1..K]$.

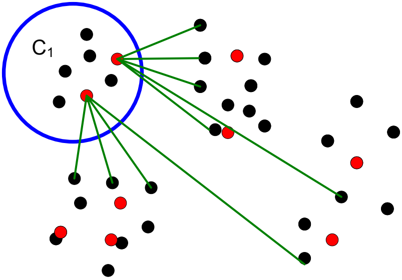

Figure: Distance Estimation in RN-partition

Partitioning-based Reference Node Embedding

Denote the closest reference node $r \in R_i$ to v as $r_{v,i}$:

$$r_{v,i} = \arg\min_{r \in R_i} D(r, v)$$

and thus

$$D(SN_i, v) = D(r_{v,i}, v) = \min_{r \in R_i} D(r, v)$$

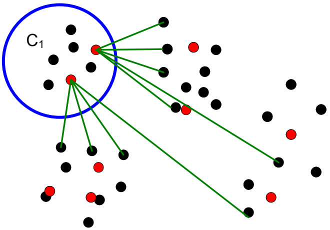

Figure: Distance Estimation in RN-partition

Partitioning-based Shortest Distance Estimation

The approximate shortest distance is estimated by

$$\widehat{D^P}(s,t) = \min_{i \in [1,K]} (D(s, SN_i) + D(r_{s,i}, r_{t,i}) + D(t, SN_i))$$

$$D(s,t) \leq \widehat{D^P}(s,t)$$

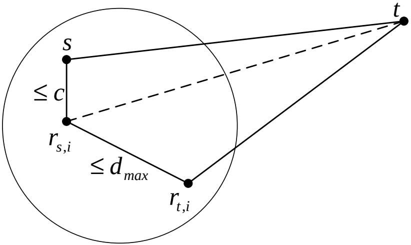

Figure: Distance Estimation in RN-partition

Partitioning-based Shortest Distance Estimation

The approximate shortest distance is estimated by

$$\widehat{D^P}(s,t) = \min_{i \in [1,K]} (D(s, SN_i) + D(r_{s,i}, r_{t,i}) + D(t, SN_i))$$

$$D(s,t) \leq \widehat{D^P}(s,t)$$

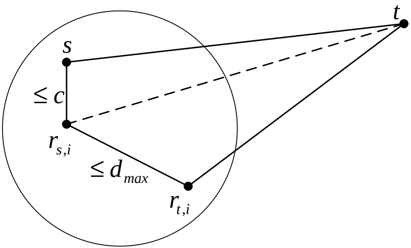

Figure: Distance Estimation in RN-partition

Definition (Cluster Diameter)

Given a cluster C, we define the diameter d of cluster C as

$$d = \max_{r_i, r_j \in C} D(r_i, r_j)$$

Error bound Analysis

Theorem

Given any query (s, t), *let* $err(s, t) = |\widehat{D^P}(s, t) - D(s, t)|$,

$$P(err(s, t) \leq 2(c + d_{max})) \geq 1 - (1 - CR)^2.$$

Complexity Comparison between RN-basic, and RN-partition

Table: Comparison between RN-basic and RN-partition

Complexity	**RN-basic**	**RN-partition**
Offline Time	$O(\|R\|n \log n)$	$O(Kn \log n + \|R\|n/K \log n/K)$
Offline Space	$O(\|R\|n)$	$O(Kn + \|R\|^2/K)$
Distance Query	$O(\|R\|)$	$O(K)$
Error Bound	$2c$	$2(c + d_{max})$

- Problem Statement
- Basic algorithms
- Graph Partitioning-based Heuristic
- Experiments
  - Case Study 1: Road Network(New York), $|V| = 264,346$, $|E| = 733,846$.
  - Case Study 2: Social Network(DBLP), $|V| = 629,143$, $|E| = 4,763,500$.

Comparison Methods and Evaluation

We compare our methods RN-basic and RN-partition with two existing methods:

- **2RNE** [5] by Kriegel et al. , and we set parameter $K = 3$.
- **Centrality** [7] by Potamias et al.

Comparison Methods and Evaluation

We compare our methods RN-basic and RN-partition with two existing methods:

- **2RNE** [5] by Kriegel et al. , and we set parameter $K = 3$.
- **Centrality** [7] by Potamias et al.

For a node pair (s, t)

$$rel\_err(s,t) = \frac{|\widehat{D}(s,t) - D(s,t)|}{D(s,t)}$$

The queries are randomly selected with size $10,000$.

Case Study 1: Road Network

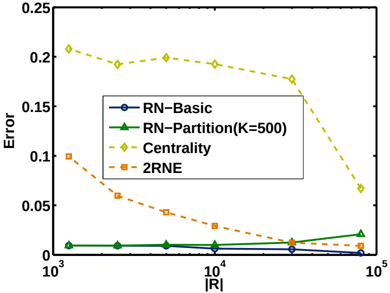

Figure: Average Error vs. $|R|$ on Road Network

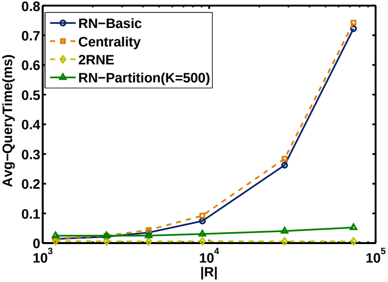

Figure: Average Query Time vs. $|R|$ on Road Network

Case Study 2: Social Network

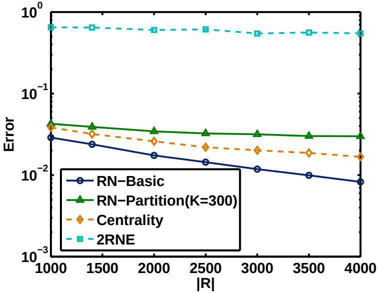

Figure: Average Error vs. $|R|$ on Social Network

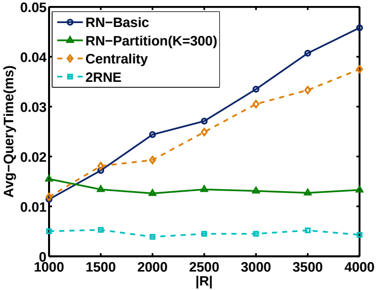

Figure: Average Query Time vs. $|R|$ on Social Network

Conclusions

Conclusion

Conclusions

Conclusion

Q&A

Conclusions

Conclusion

Q&A

Thanks!

P. Francis, S. Jamin, C. Jin, Y. Jin, D. Raz, Y. Shavitt, and L. Zhang.
IDMaps: A global internet host distance estimation service.
IEEE/ACM Trans. Networking, 9(5):525–540, 2001.

A. Gubichev, S. Bedathur, S. Seufert, and G. Weikum.
Fast and accurate estimation of shortest paths in large graphs.
In *Proc. 2010 Int. Conf. Information and Knowledge Management (CIKM'10)*, 2010.

G. Karypis and V. Kumar.
A fast and high quality multilevel scheme for partitioning irregular graphs.
SIAM Journal on Scientific Computing, 20(1):359–392, 1999.

S. Khuller, A. Moss, and J. Naor.
The budgeted maximum coverage problem.
Information Processing Letters, 70:39–45, 1999.

H.-P. Kriegel, P. Kröger, M. Renz, and T. Schmidt.
Hierarchical graph embedding for efficient query processing in very large traffic networks.
In *Proc. 2008 Int. Conf. Scientific and Statistical Database Management (SSDBM'08)*, pages 150–167, 2008.

T. S. E. Ng and H. Zhang.
Predicting internet network distance with coordinates-based approaches.
In *Int. Conf. on Computer Communications (INFOCOM'01)*, pages 170–179, 2001.

M. Potamias, F. Bonchi, C. Castillo, and A. Gionis.
Fast shortest path distance estimation in large networks.
In *Proc. 2009 Int. Conf. Information and Knowledge Management (CIKM'09)*, pages 867–876, 2009.

M. J. Rattigan, M. Maier, and D. Jensen.

Using structure indices for efficient approximation of network properties.
In *Proc. 2006 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'06)*, pages 357–366, 2006.

A. D. Sarma, S. Gollapudi, M. Najork, and R. Panigrahy.
A sketch-based distance oracle for web-scale graphs.
In *Proc. 2010 Int. Conf. Web Search and Data Mining*, pages 401–410, 2010.

C. Shahabi, M. Kolahdouzan, and M. Sharifzadeh.
A road network embedding technique for k-nearest neighbor search in moving object databases.
In *Proc. 10th ACM Int. Symp. Advances in Geographic Information Systems (GIS'02)*, pages 94–100, 2002.

M. Thorup and U. Zwick.
Approximate distance oracles.
Journal of the ACM, 52(1):1–24, 2005.