Complete Graph Reconstruction from Partial Information

Gunjan S. Mahindre and Anura P. Jayasumana

gunjanm5@colostate.edu   anura@engr.colostate.edu

Electrical and Computer Engineering, Colorado state University, Fort Colling, CO 80523

Research Question: If we only know the distances from a few nodes in a graph, how do we reconstruct the original graph?

Abstract

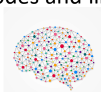
Networks, biological molecules, neural structures can be represented as graphs. Data processing and storage of such structures with millions of nodes is very bulky.

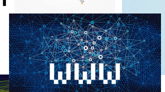
Thus, we derive important properties and synthesize a technique to regenerate a graph from partial information about the graph with minimum data and high fidelity.

This will impact the way we store and operate on network data and opens new possibilities in areas such as chemistry, social networks, neural networks and the ever evolving Internet.

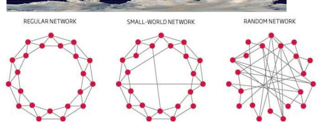
Motivation

- Graphs with millions of nodes and links
- **Social networks**
- **World Wide Web**
- **Neural networks in Brain**
- **Airline networks**
- Generates Big Data

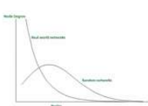




- Real world networks
- Distributed systems require compact data representation



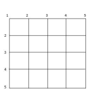
Degree distribution of nodes



Compact representation of large graphs with zero or minimal loss of Information

Graphs as Data

- Link information & distance information
- Adjacency and distance matrices represent graphs



Links
$$\begin{bmatrix} 0 & 1 & 0 & .. & 0 \\ 1 & 0 & 1 & .. & 0 \\ 0 & 1 & 0 & .. & 0 \\ .. & .. & .. & .. & .. \\ 0 & 0 & 0 & .. & 0 \end{bmatrix}$$

Distances
$$\begin{bmatrix} \boxed{0} & 1 & \boxed{2} & .. & 8 \\ \boxed{1} & 0 & \boxed{1} & .. & 7 \\ \boxed{2} & 1 & \boxed{0} & .. & 6 \\ \boxed{..} & .. & \boxed{..} & .. & .. \\ \boxed{8} & 7 & \boxed{6} & .. & 0 \end{bmatrix}$$

- Large graphs generate huge data: $[\boldsymbol{N} \times \boldsymbol{N}]$
- Information in $[\boldsymbol{N} \times \boldsymbol{N}]$ ~ Information in $[\boldsymbol{N} \times \boldsymbol{M}]$; $\boldsymbol{M} \ll \boldsymbol{N}$

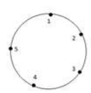
How to select those $\boldsymbol{M}$ nodes to get $[\boldsymbol{N} \times \boldsymbol{N}]$ information?

Prior Work

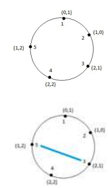
- **Metric Dimension** - $\beta(\boldsymbol{G})$: minimum number of anchor nodes that create unique distance vectors
- **Resolution set** - $\{\boldsymbol{R}\}$: collection of these anchor nodes

Challenge

Metric dimension is not always enough



$$\begin{bmatrix} \boxed{0} & \boxed{1} & 2 & 2 & 1 \\ \boxed{1} & \boxed{0} & 1 & 2 & 2 \\ \boxed{2} & \boxed{1} & 0 & 1 & 2 \\ \boxed{2} & \boxed{2} & 1 & 0 & 1 \\ \boxed{1} & \boxed{2} & 2 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 2 & 1 \\ 2 & 2 \\ 1 & 2 \end{bmatrix}$$



Original Graph | Complete Distance Matrix | Selected - Distance vectors | Ambiguous Reconstruction

Ambiguities in reconstruction

Our Goal: Unique and complete graph reconstruction

Contribution

- Predict presence of a link
- **Theorem 1: Cause of ambiguity** - An edge between nodes i and j is invisible if and only if for each anchor $\boldsymbol{A}_k$,

hop distance between nodes i and A_k

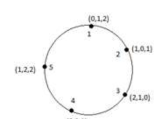
$$\{Max|\, h_{iA_k} - h_{jA_k} \,| \leq 1\}$$
$$\forall\, i,j \in \mathcal{N};\ \forall\, k;\ \boldsymbol{A}_k \in \mathcal{A}, k = 1{:}\,\boldsymbol{M}$$

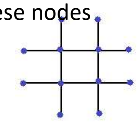
- Eliminate ambiguity with additional well placed anchor nodes
- **Theorem 2:** The Link dimension, metric dimension, resolution set and reconstruction set can be related as:

Link Dimension · *Cardinality of Reconstruction set* · *Cardinality of Resolution set* · *Metric Dimension*

$$\{\alpha(\boldsymbol{G}) \geq |\{\boldsymbol{C}\}| \geq |\{\boldsymbol{R}\}| \geq \beta(\boldsymbol{G})\}$$

- **Link Dimension** - $\alpha(\boldsymbol{G})$: minimum number of nodes to predict all links without any ambiguity
- **Reconstruction set** - $\{\boldsymbol{R}\}$ collection of these nodes





Metric dimension = 3

Link Dimension = 6

Data compression ratio = 1/2
= Required data/ Actual data
= 6x12 matrix / 12x12 matrix
= ½

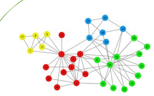
Procedure

- **Finding nodes in Reconstruction set:**

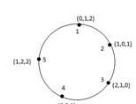
```
INPUT: distance matrix, Resolution set {R}
OUTPUT: {C} reconstruction set
starting from {R}, where |{R}| = β(G),
Let |h_iAk − h_jAk| = Δ
DO
//record node pairs with ambiguity
pairs = []
FOR i = 1: N
    FOR j = 1: N
        IF Δ> 1
            append (i, j)to pairs
        END
    END
END
FOR n = 1: N
  FOR each ambiguity pair (i, j)
     IF { |dist_ki − dist_kj| > 1 }; i, j and k ∈ G(V)
        // node n resolves ambiguity
          Strength[n] ++    // number of ambiguities resolved
     END
END
Select k s.t. Strength[k] = max(Strength)
Remove the ambiguity pairs resolved by node k
WHILE (ambiguity>0)
```

- **Other findings:**
  - Reliability of a given Distance Vector
  - Validity of given Distance matrix
  - Compute bounds for missing distances
- **Challenges:**
  - Optimum selection from a large pool of nodes
  - Predicting anchor nodes
  - Finding minimum solution

Future Work



- Web spam detection
- Complete graph reconstruction
- Integrate machine learning to recognize silent characteristics of a graph



$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \\ 2 & 2 & 1 \\ 1 & 2 & 2 \end{bmatrix}$$

- Data compression
- Community detection
- Performing distributed functions with lower computational cost
- An efficient algorithm for any graph to determine
  - Number and location of minimum constructors
  - Extend work for directed graphs