

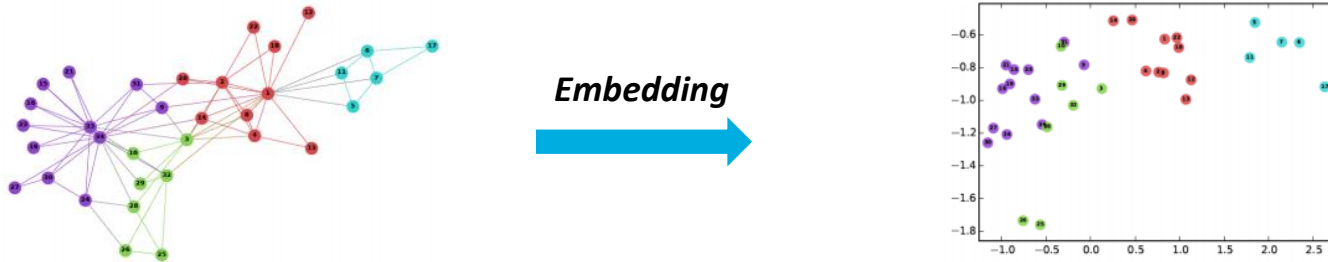
Integrate Link Inference and Deep Graph Convolutional Network (LIGCN)

Sheng Guan Hanchao Ma Chen Guo Yunzhou Cao



Introduction (Node Embedding)

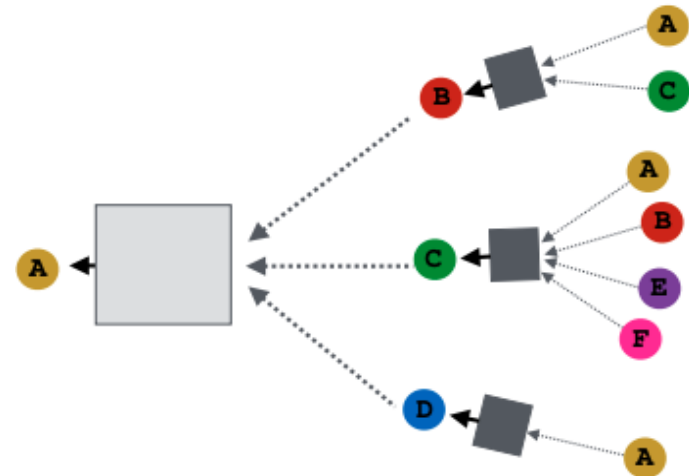
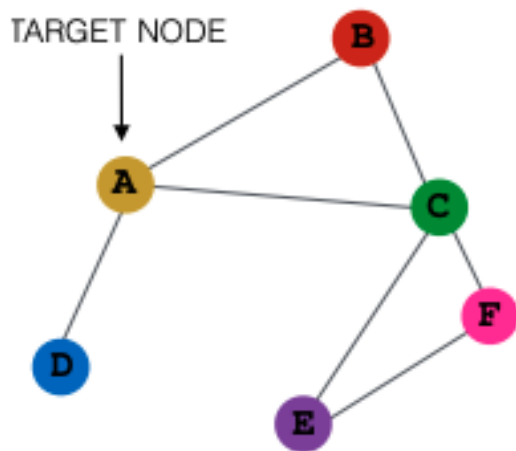
- **Intuition:** Embedding of nodes to d-dimensions(low) so that “similar” nodes in the graph have embeddings that are close together.

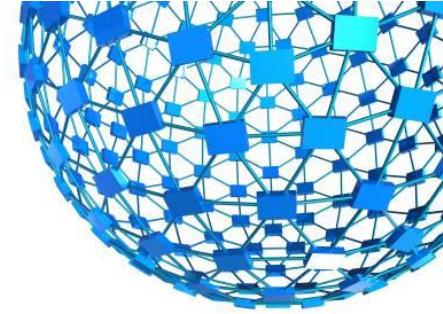


- **Goal:** Encoding nodes so that similarity in the *embedding* space (e.g., dot product) approximates similarity in the *original network*.
- **Challenges:**
 - How to define encoder:
 - Node2Vec
 - Random Walk
 - How to define similarity function.(connected? , share attributes?)
 - Adjacency-based Similarity
 - Multi-hop Similarity

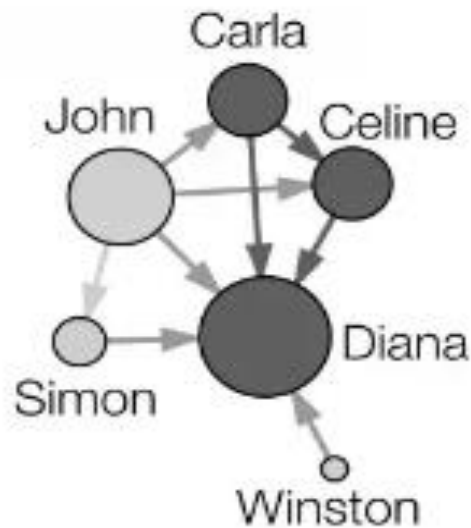
Graph Neural Networks

- **Why GCN:** Traditional approaches have limitations.
 - $O(|V|)$ parameters are needed (no sharing parameters)
 - Can not generate embeddings for nodes that were not seen during training
 - Do not incorporate node features
- **Key Idea:** Generate node embeddings based on local neighborhoods.
- **Example of an embedding (boxes are the model we want to train)**





Friends Recommendation In Social Network



Nodes

Id,Label,Attribute

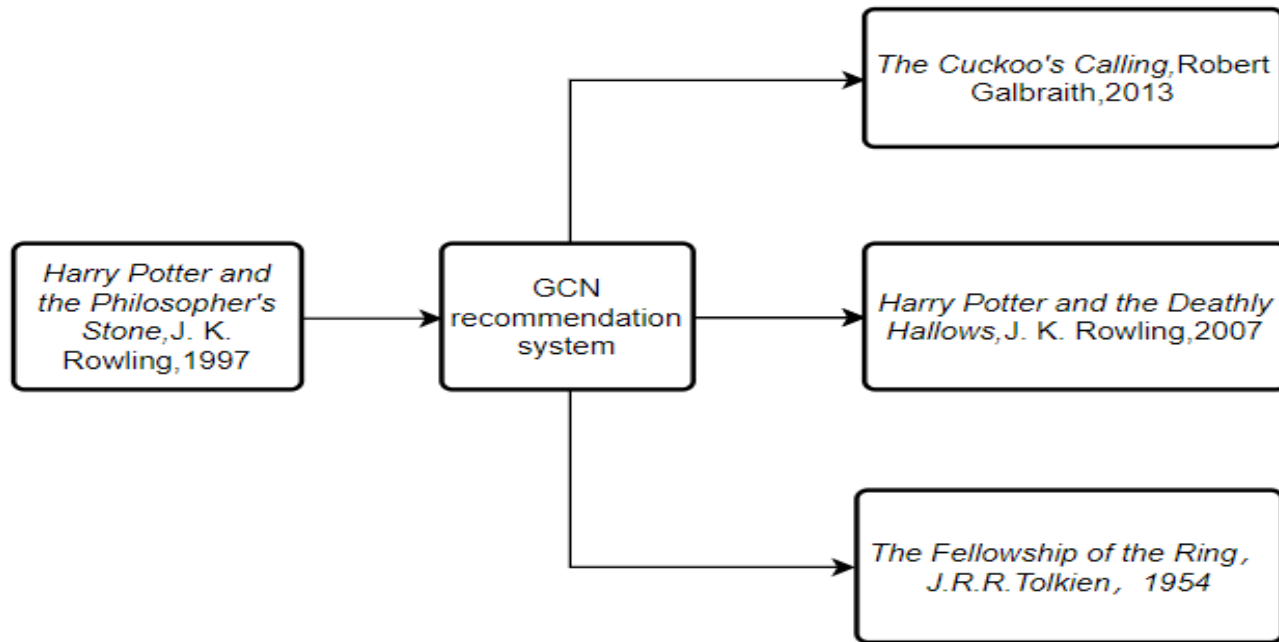
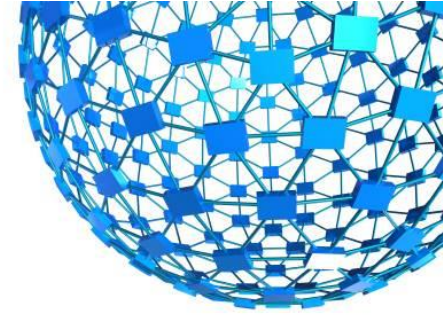
1,John,1
2,Carla,2
3,Simon,1
4,Celine,2
5,Winston,1
6,Diana,2

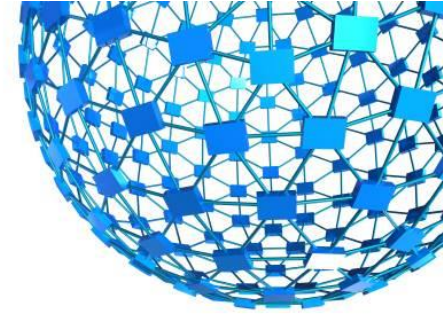
Edges

Source,Target

1,2
1,3
1,4
1,6
2,4
2,6
3,6
4,6
5,6

Book Recommendation System



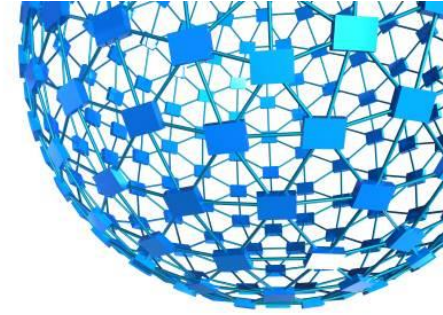


Challenges

- Even for the GCN, it still assumes the training data is **100% clean**, but that's not the case in real life
(graph can be **incomplete and erroneous**)
- Stacking multiple graph convolutional layers over-smooth features and limits the model with a shallow structure
(cannot capture **long-range** dependencies!!!)
- There is **no** graph neural network approach targeting erroneous entities detection in attributed networks

GCN still has limitations and not for erroneous entities detection

Problem Definition (Book Recommendation)

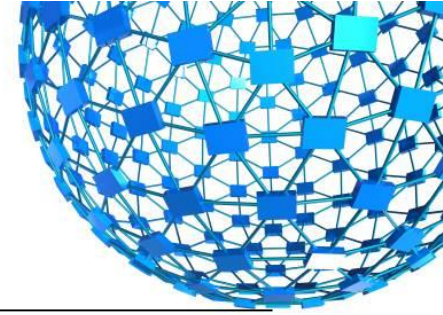


Book recommendation problem.

Having trouble deciding which one to read among the countless books on the Internet, a better recommendation system always benefits customers and the book providers.

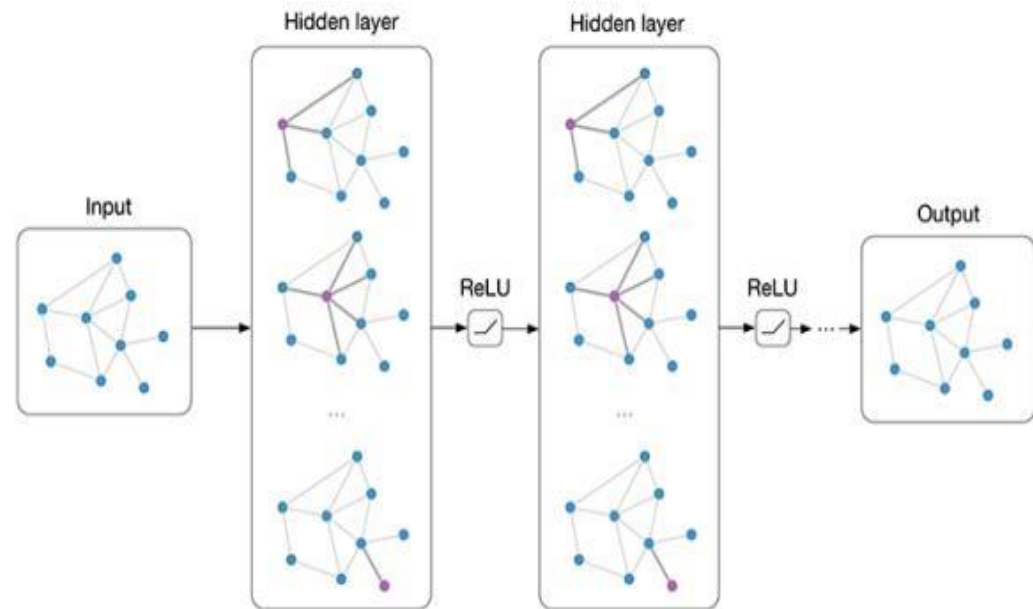
What's more, we believe Graph Convolution Network(GCN), which is relatively new and has a promising future approach can solve problems related to irregular graph.

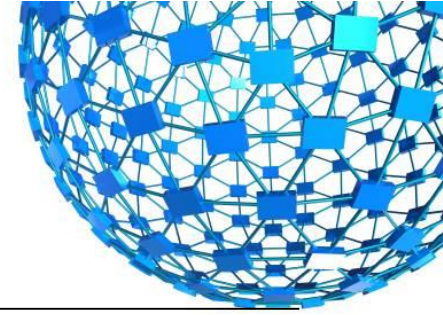
Approach (Book Recommendation)



Embed each node's attributes (e.g. Book-Title, Book-Author, Year-Of-Publication, Publisher) into vectors.

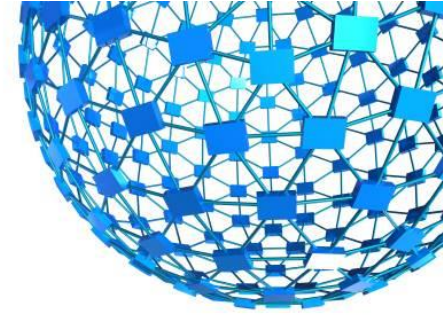
Network structure







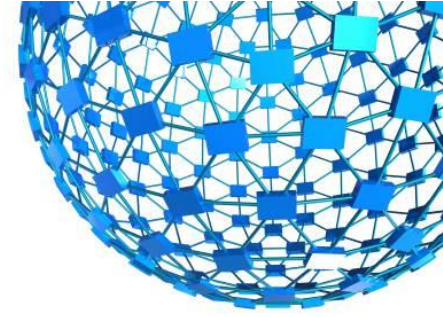
Problem Definition (LIGCN)

- Input: An attributed network $G = (V, E, X)$ consists of:
 - (1) the set of nodes $V = \{v_1, v_2, \dots, v_n\}$; $|V| = n$
 - (2) the set of edges; $|E| = m$
 - (3) the node attributes X
 - in the semi-supervised setting, we only have limited labels on partial nodes that indicate whether the node is clean or contains dirty information.
- Output: evaluate the performance of all testing labeled nodes to achieve better performance, e.g. accuracy, F-1 score, etc.



Approach

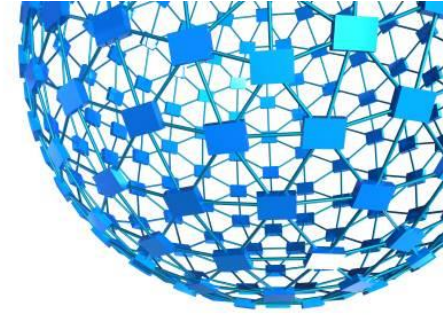
- To actively identify potential n-hop neighbors that contribute to the detection of local erroneous information at node v
supervised random walk  learn “hidden link”
- A novel GCN-based deep graph autoencoder
detect erroneous entities by measuring the reconstruction errors of node attributes  enhance error detection



- Related Work (Recommender System Case Study)

We search for suitable dataset for this very application.

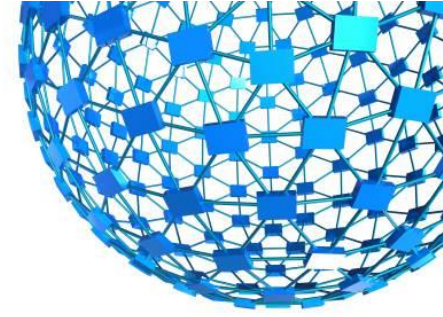
We read some related research papers, e.g., R. Ying et al. Graph convolutional neural networks for web-scale recommender systems. KDD, 2018.



Related Work (**LIGCN**)

- ❖ Rule-based Model for Erroneous Entity Detection
 - Assume graph is complete, incomplete graph will lead to **incomplete set of rules**
 - Need to compute pattern matching and are **computationally hard** in general

LIGCN doesn't need to build rules from the scratch

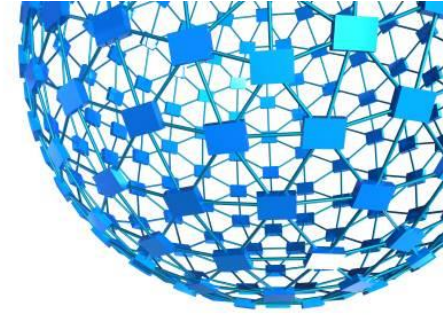


Related Work (**LIGCN**)

- ❖ Learning-based Model for Erroneous Entity Detection
 - LIGCN shows the significance of developing a novel deep architecture that can **better model the non-linearity** between the node interactions and nodal attributes
 - LIGCN is the first attempt to consider **combining graph completion and embedding learning** together

LIGCN builds better deep architecture

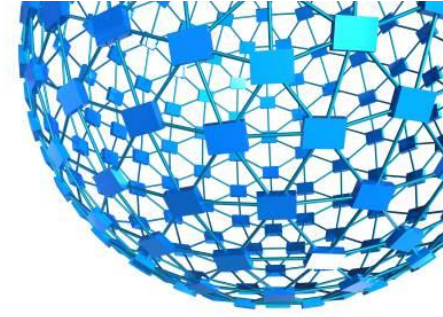
Recommendation (dataset)



Book-Crossing dataset will be used in the current work. Each book is represented as itself ISBN, title, author (only first one), year of publication and publisher.

Data set is obtained at <http://www2.informatik.uni-freiburg.de/~cziegler/BX/>

Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, Georg Lausen; *Proceedings of the 14th International World Wide Web Conference (WWW '05)*, May 10-14, 2005, Chiba, Japan.



LIGCN (dataset)

Synthetic dataset: BSBM (e-commerce benchmark)

Able to generate synthetic knowledge graph over a set of products and orders

Controlled by # of nodes (up to 60M)

of edges (up to 152M)

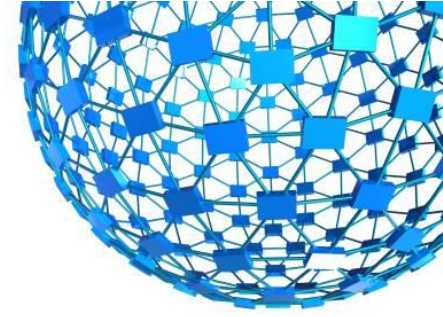
of attribute labels (up to 3080)

Need pre-processing:

Aggregate related attributes and further get features

Generate a connected graph

LIGCN (dataset)



Real-world graph datasets:

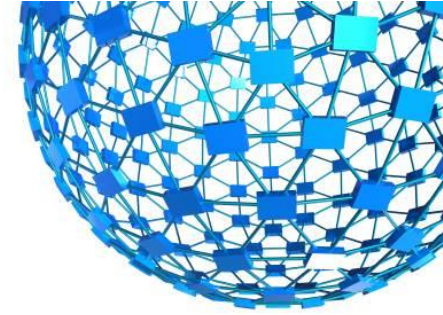
- Weibo (social network)

- DBpedia (Knowledge graph)

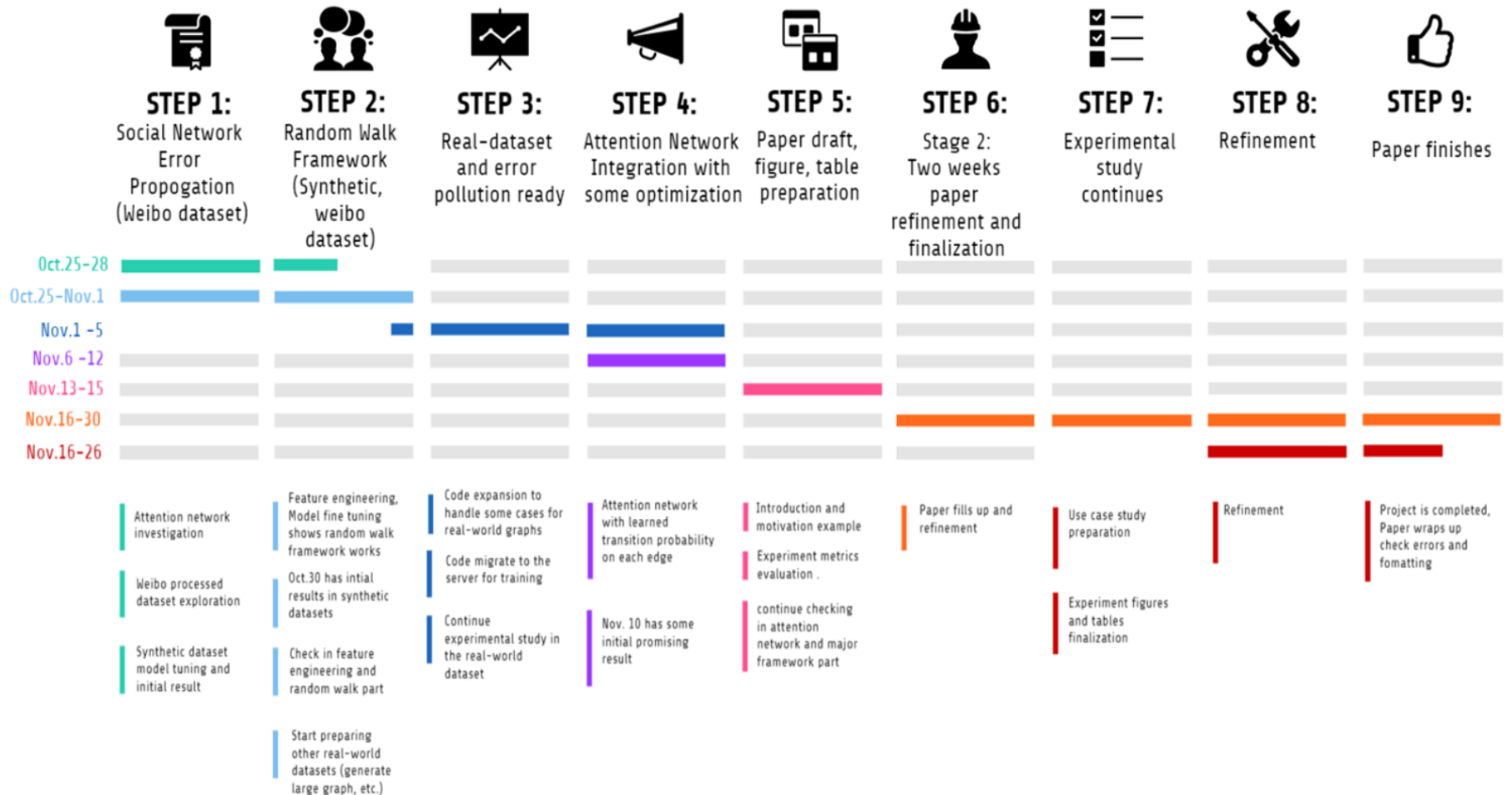
Generate errors:

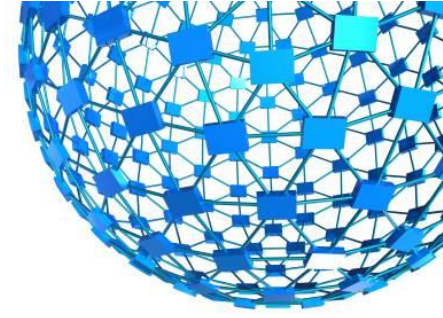
For Weibo dataset, we adopt propagation model (e.g. rumor propagation) to simulate error generation

For DBpedia dataset, we adopt BART framework to generate errors according to functional dependencies and random error generation



Project Implementation

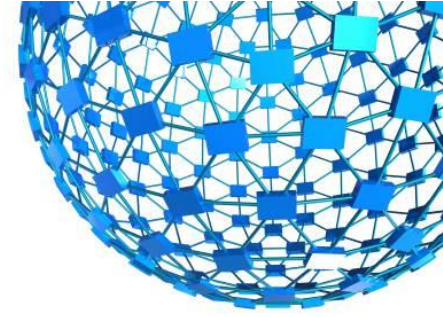




Evaluation Metrics(Book recommendation)

The evaluation metrics are mainly three parts:

- Accuracy
- Loss value
- Volunteer customers' evaluation



Evaluation Metrics and baselines (**LIGCN**)

We can treat the task of LIGCN as classification problem, then the evaluation metrics:

- Accuracy
- Precision/ Recall/ F-1 score
- ROC curve with AUC area

Baselines:

- GCN
- GraphSAGE (a variant of GCN with more complex aggregation function)

Thank you very much!



Backup Slides

- Questions you should answer in your presentation –
 - What is the problem you are addressing and why is it pertinent?
 - What research have you done into the context of this problem?
 - What is the data that you will use? How did/will you get it?
 - How are you planning on implementing your solution?
 - What metrics are you using to evaluate your results?