Modelling hackforum networks using node2vec

Student Name: Jash Shah and Akash Pratap Singh
Meeting time: Every Tuesday, 12:00hrs
Meeting Link: https://meet.google.com/nhx-rvms-erm
Mobile contact: 9660767800(Akash) | 9825846159(Jash)
Email - (Jash Shah) f20180507@pilani.bits-pilani.ac.in

- (Akash) f20180462@pilani.bits-pilani.ac.in

Task:

1. Implementation of node2vec on Karate club network
2. Understanding functionality of each function used
3. Study of other random walk based algorithms

Sources

1. Node2vec code
   https://github.com/aditya-grover/node2vec
2. Node2vec article
   https://snap.stanford.edu/node2vec/
3. Maithili Sharan Gupta IPS, interaction with SIH Teams , PSID- MS332: Textile recognition for characterizing criminals
   https://www.youtube.com/watch?v=A9P5WZyP4qM

Please update the meeting minutes after the meeting and progress report one day before the meeting.

Week 30:

Meeting Minutes:

- Install Python, Run node2vec data on graph data available in karate.edgelist.
- Unique Function, use for undirected graphs.
- Run main() function which calls node2vec(), give proper args & save in .emb files.
- Installation of setup for .sql files, for hackforum database..
- Try to understand where BFS & DFS strategies are used.

Progress report:

- Resolved all errors & finally ran main.py successfully with python3, verified results with author's experiment.
- Studied python syntax & Numpy module.
- Completed server setup & generated table using phpMyAdmin for a 200kb sample database sql file. Problem encountered, It has a limit of 40 MB.

Week 31:

Meeting Minutes:

- Discussion of how to scrape dark web data.
- Setting up a secure tunnel to access the dark web.
- Explanation of Hackforum.
- Demonstration of setting up Community SQL server.
- Exploration of Ironmarch database
- Discussion of applying node2vec in Ironmarch database.

Progress report:

- Converting SQL data in graph format.
- Run node2vec on the available database.
- Completed : Extraction of relevant data from ironmarch database & converted into an edgelist file.
- Embeddings generated after running node2vec are available in ironmarch.emd file.
- Github Link to generated embeddings :- https://github.com/Apsakash1/node2vec-results-on-ironmarch-nulled-data/tree/master/node2vec-master/emb

Week 32:

Meeting Minutes:

- Presentation of understanding of node2vec.
- Suggestion to think of an approach to generate weights for the data from orig_topics.
- Figure out the type of task we will need to perform once we get the data in Indian Context.
- Explore other algorithms which can be used.

Progress report:

- Node2vec performs better on node classification while multi-hop methods perform better on link prediction(Goyal and Ferrara, 2017 Survey).
- Generate embeddings of dimensions nx1, then use these to input as weights to the node2vec with the node connections.
- Compare the results from previous results to see which is able to represent the information more closely.
- Suggestions on how to proceed further :-
    1) Use any permutation or combination of embeddings.
    2) Use PCA or dimensionality reduction to reduce in 1 or 2 or 3 dimensions.
    3) Use random number generation.
    4) Label some of the nodes 0 & 1 where indicates potential threat level maximum. Use this small labeled set of data & input to classification on GCP(codeless) or use any other way to classify the complete set of data. Then use the confidence level value as weight(Say ML algo's gives output ! for a specific edge with 82% confidence then use .82 as it's weight).

Week 33:

Meeting Minutes:

- Brief overview of Presentation for understanding of node2vec.
- Presentation of Ideas to generate weights.
- Discussion of using euclidean distance to use 128 dimensional vectors after normalization as edge weight to the graph.
- Understand node2vec to modify the algorithm to guide our walks to explore different network neighbourhoods. OR Implement a similar algorithm which can be modified to generate results.

Progress report:

- Understood the functionality of all the functions used and now I can modify them to explore the network accordingly.
- Implemented node2vec from scratch for in depth understanding.
- Understanding & Implementation of Sampling strategy with O(1) time complexity.
- Using tensorboard embedding projector Visualization of embeddings generated.
- Visualization of embeddings generated.

- Script for conversion of node2vec embeddings to tensors and metadata file for use in embedding projector.
- Completed the presentation for mid sem seminar.

Week 34:

Meeting Minutes:

- Midsem Seminar.
- Presentation of all functions used in node2vec implementation in detail.
- Preprocessing of Transition probabilities.
- Aliasing function to make the next step of walk.
- Discussion of word2vec
- Recursive algorithm.
- Data Visualisation techniques using tensorflow embedding projector.
- Results of tSNE for clustering even if it's a dimensionality reduction algorithm.

Progress report:

- Aliasing function link :- https://lips.cs.princeton.edu/the-alias-method-efficient-sampling-with-many-discrete-outcomes/
- Presentation (PDF format) - https://mail.google.com/mail/u/0?ui=2&ik=33df96aaec&attid=0.2&permmsgid=msg-a:r-1305241341096658191&view=att&disp=safe&realattid=f_kfnr0tt71

Week 35:

Meeting Minutes:

- Midsem Seminar continued.
- Revisiting & Clarification of node2vec implementation in detail.
- Biased random walk definition improvisation suggested.
- Revisited Aliasing function to calculate the next step of walk.
- Discussion of word2vec objective.
- Show Results of tSNE for clustering for Karate club and IronMarch data and Karate club data.
- Discussion on the concept of similarity and diversity instead of p & q parameters.

Progress report:

- Script for conversion of embeddings to visualise in embedding projector.
- Think of a way to make our random walk to consider the previous two steps into consideration instead of one while calculating the next step.
- Try to think of an approach for including similarity and diversity concepts for biasing random walks. This will make the process of completely automatic instead of fixed values set manually and it will be a more appropriate strategy.

- Google drive Link :- https://sgnldrp.online/click?redirect=https%3A%2F%2Fdrive.google.com%2Fdrive%2Ffolders%2F1cv4dzRZJKdLFEpQfFrRQpOH9FAA7aOu-%3Fusp%3Dsharing&dID=1602056425284&linkName=https://drive.google.com/drive/folders/1cv4dzRZJKdLFEpQfFrRQpOH9FAA7aOu-?usp=sharing

Week 36:

Meeting Minutes:

- Task - I : use facebook data and try to replicate the results obtained in the paper for link prediction in the similar fashion.
- Task - II : Use concept of similarity instead of fixed value of q for BFS & DFS.
- Bridge/gateway detection presentation.
- Discussion of sub2vec & other algorithms.

Progress report:

- Verified the efficiency claims made by authors by applying node2vec on Facebook Ego networks for Link prediction.
- Explored the concept of similarity instead of fixed value of q.
- Studied Random walk based algorithms other than node2vec like Deepwalk.
- Prepared report & End Sem presentation along with the future plans of working on the project.